Financiada por el MICIIN
(TSI2007-30967-E)

Jornadas PLN-TIMM, 5 y 6 de febrero de 2009
Organizadas por la red TIMM*

Campus de Colmenarejo, Universidad Carlos III de Madrid

III JORNADAS SOBRE MODELOS Y TÉCNICAS PARA EL ACCESO A LA INFORMACIÓN MULTILINGÜE Y MULTIMODAL EN LA WEB

Actas de las ponencias

* Cofinanciadas por la red MAVIR (www.mavir.net)

ÍNDICE

Emociones, opiniones e identidad

Semántica y sintáctica

Búsqueda y navegación

Interacción y ontologías para el acceso a servicios

Edición de las actas: Ana García Serrano (UNED) y Cesar de Pablo (UC3M)

PRESENTACIÓN

La red temática TIMM (Tratamiento de Información Multilingüe y Multimodal) dentro del programa de acciones complementarias (TSI2007-30967-E) http://sinai.ujaen.es/timm/, da soporte tanto a las III jornadas sobre "Modelos y técnicas para el acceso a la información multilingüe y multimodal en la web", como a su organización en el Campus de Colmenarejo del Departamento de Informática de la Universidad Carlos III de Madrid.

El objetivo de las jornadas es promover la difusión de las actividades de investigación, desarrollo e innovación entre los diferentes grupos de investigación. Se persiguen dos objetivos específicos:

- crear un foro donde los investigadores en formación puedan presentar y discutir su trabajo en un ambiente que facilite el intercambio de ideas y la colaboración.
- realizar un catálogo de recursos lingüísticos y herramientas desarrolladas en los diferentes grupos de investigación para fomentar su uso y difusión entre otros grupos.

En esta memoria de actas, se incluyen los resúmenes y descripciones breves de las ponencias presentadas.

Todos los organizadores de las jornadas deseamos expresar nuestro agradecimiento al Departamento de Informática de la UC3M y en concreto al grupo de investigación LABDA, dirigido por la Profesora Dña. Paloma Martínez, por su soporte y apoyo a las mismas. Finalmente agradecer la cofinanciación al consorcio MAVIR (http://www.mavir.net) e indudablemente a la red TIMM, en cuyo marco se organiza por tercera vez estas jornadas.

Los organizadores de las III jornadas PLN-TIMM:

Cesar de Pablo, UC3M (cdepablo@inf.uc3m.es

Ana García Serrano, UNED (agracia@lsi.uned.es)

Maite Martín, Universidad de Jaén (maite@ujaen.es)

Víctor Peinado, UNED (victor@lsi.uned.es)

L. Alfonso Ureña, Universidad de Jaén (laurena@ujaen.es)

Emociones, opiniones e identidad

Pros and Cons: Sentiment Analysis Applied to Multilingual, Multi-genre Texts

Alexandra Balahur, Andrés Montoyo

Departamento de Lenguajes y Sistemas Informaticos, Universidad de Alicante
Apartado de Correos 99, E-03080, Alicante, España
{abalahur, montoyo}@dlsi.ua.es

Abstract. Sentiment analysis (opinion mining) is a difficult task due to the high semantic variability of natural language. It supposes not only the discovery of directly expressed opinions, but also the extraction of phrases that indirectly or implicitly value objects, by means of emotions or attitudes. In this article we present the contributions we brought to the field of sentiment analysis. They reside in the creation of a lexical database of terms for emotion detection, which we denoted "emotion triggers", the proposal of several methods for feature-based opinion mining, the application and evaluation of the opinion mining methods to product review summarization, as component of a recommender system, of a multi-perspective question answering system and for opinion tracing over political debates. Each of the resources and methods proposed were evaluated and showed good and promising results.

Keywords: opinion mining, sentiment analysis, emotion detection, polarity classification.

1 Introduction and Motivation

Recent years have marked the beginning and expansion of the social web, in which people freely express and respond to opinion on a whole variety of topics. The growing volume of opinion information available allows for better and more informed decisions of the users, the quantity of data to be analyzed imposed the development of specialized Natural Language Processing systems that automatically extract, classify and summarize the opinions available on different topics. Research in this field, of opinion mining (sentiment analysis), has shown opinion mining is a difficult problem and addressed it from different perspectives and at different levels, depending a series of factors. These factors include: *level of interest* (overall/specific), *querying formula* ("Nokia E65"/"Why do people buy Nokia E65?"), *type of text* (review on forum/blog/dialogue/press article), and *manner of expression of opinion* - directly (using opinion statements, e.g. "I think this product is wonderful!"/"This is a bright initiative"), indirectly (using affect vocabulary, e.g. "I love the pictures this camera takes!"/"Personally, I am shocked one can propose such a law!") or implicitly (using

adjectives and evaluative expressions, e.g. “It’s light as a feather and fits right into my pocket!”).

The aim of our work has been to create, exploit and test both new and consecrated resources that help to detect emotion, opinion and attitude and subsequently classify them according to their polarity (positive/negative).

2 Emotion Triggers

The aim in building this resource is to draw the attention upon the difference between the *cognitive* and *emotional* aspects of text, as theoretically explained by the *Theory of Emotivism* [17] and investigate a method to obtain a database of terms that being related to human needs and motivations. We define a new concept, called “emotion trigger”, whose definition can be summarized by: “An **“emotion trigger”** is a word or idea that is connected to general human needs and motivations or that depending on the reader’s interests, cultural, educational and social factors, relates to general human needs and motivations and thus leads to an emotional interpretation of a given text. (e.g. war, freedom, mother, bomb)” [2]. The approach is theoretically underpinned by the Theory of relevance [16] – giving different importance of emotion triggers according to “relevance”. The core of “emotion triggers” is taken from Abraham Maslow’s Pyramid of human motivations [14] and from Manfred Max-Neef’s matrix of human needs and satisfiers [15]. The first one contains the general human motivations in a hierarchy of 5 levels, from the 3 bottom levels that are basic, to the upper 2 levels containing the higher needs. Max-Neef’s matrix contains terms organized according to 4 existential categories and 9 categories of needs (therefore having different relevance). The core of terms is expanded using lexical resources such as WordNet and WordNet Affect, completed by NomLex, sense number disambiguated using the Relevant Domains concept. The mapping among languages is accomplished using EuroWordNet and the completion and projection to different cultures is done through language-specific commonsense knowledgebases. Subsequently, we show the manner in which the constructed database can be used to mine texts for valence (polarity) and affective meaning. An evaluation is performed on the Semeval Task No. 14, obtaining better results than the systems participating in the competition. For further details, please see [1,2].

3 Opinion Mining System

Our system is based upon the feature-based opinion mining paradigm. For each product class we first automatically extract general features (characteristics describing any product, such as price, size, design), for each product we then extract specific features (as picture resolution in the case of a digital camera) and feature attributes (adjectives grading the characteristics, as for example high or low for price, small or big for size and modern or faddy for design). Further on, we assign a polarity

(positive or negative) to each of the feature attributes using a previously annotated corpus and Support Vector Machines Sequential Minimal Optimization [10] machine learning with the Normalized Google Distance [11] and Latent Semantic Analysis [12], as well as patterns of affect expressions. We show how the method presented is employed to build a feature-driven opinion summarization system that is presently working in English and Spanish. In order to detect the product category, we use a modified system for person names classification. The raw review text is split into sentences and depending on the product class detected, only the phrases containing the specific product features are selected for further processing. The phrases extracted undergo a process of anaphora resolution, Named Entity Recognition and syntactic parsing. Applying syntactic dependency and part of speech patterns, we extract pairs containing the feature and the polarity of the feature attribute the customer associates to the feature in the review. Eventually, we statistically summarize the polarity of the opinions different customers expressed about the product on the web as percentages of positive and negative opinions about each of the product features. We show the results and improvements over baseline, together and discussion on the strong and weak points of the method. For further details, please see [3, 5].

4 Recommender System

Finding the "perfect" product among the dozens of products available on the market is a difficult task for any person. Designing the "perfect" product for a given category of users is a difficult task for any company, involving extensive market studies and complex analysis. We designed a method to gather the attributes that make up the "perfect" product within a given category and for a specified community. The system built employing this method can recommend products to a user based on the similarity of the feature attributes that most users in his/her community see as positive for the product type and the products the user has to opt from and as a practical feedback for companies as to what is valued and how, for a product, within a certain community. For each product class, we first automatically extract general features (characteristics describing any product, such as price, size, and design), for each product we then extract specific features (as picture resolution in the case of a digital camera) and feature attributes (adjectives grading the characteristics, as modern or faddy for design). Further on, we use "social filtering" to automatically assign a polarity (positive or negative) to each of the feature attributes, by using a corpus of "pros and cons"-style customer reviews. Additional feature attributes are classified depending on the previously assigned polarities using Support Vector Machines Sequential Minimal Optimization [10] machine learning with the Normalized Google Distance [11]. Finally, recommendations are made by computing the cosine similarity between the vector representing the "perfect" product and the vectors corresponding to products a user could choose from. For further details, please see [4].

5 Opinion Question Answering and Summarization (TAC 2008)

The Opinion Pilot task in TAC 2008 consisted in generating summaries from opinions in the *Blog 6* collection, according to opinion questions provided by the TAC organizers on 25 targets. A set of text snippets containing the answers to these questions was also provided, their use being optional. Our system used of two methods for opinion mining and summarization, one employing the optional text snippets provided by the TAC organization (the Snippet-driven Approach) and one in which we performed the IR task as well (Blog-driven Approach). In the question processing part, we extract the topic and determine the question polarity with question patterns. These patterns take into consideration the interrogation formula and extract the opinion words (nouns, verbs, adverbs, adjectives and their determiners). The opinion words are then classified in order to determine the polarity of the question, using the WordNet Affect [8] emotion lists, the emotion triggers resource, a list of four attitudes that we built, containing the verbs, nouns, adjectives and adverbs for the categories of criticism, support, admiration and rejection and two categories of value words (good and bad) taken from the opinion mining system in [3]. In the first approach, we use the given snippets to search for the original blog sentences, extract and classify them, thus determining the answer to each corresponding question. The reformulation patterns are used to build the summary, giving coherence and structure to the texts. In the second approach, we search for possible answers directly in the blogs, determine their polarity and thus establish their correspondence to the questions they answer. Again, the summaries are generated using the reformulation patterns; since we have a 7000 character limit on each summary, we only include the strong negative and strong positive sentences, according to their similarity scores to the affect and opinion categories. For details, please see [6].

6 Opinion Tracing Across Political Debates

Finally, we investigate different approaches we developed in order to classify opinion and discover opinion sources from text, using affect, opinion and attitude lexicon. We apply these approaches on a corpus of American Congressional speech data. We propose three methods to classify opinion at the speech segment level, firstly using similarity measures to the affect, opinion and attitude lexicon, secondly dependency analysis and thirdly SVM machine learning. Further, we study the impact of taking into consideration the source of opinion and the consistency in the opinion expressed, and propose three methods to classify opinion at the speaker intervention level, showing improvements over the classification of individual text segments. Finally, we propose a method to identify the party the opinion belongs to, through the identification of specific affective and non-affective lexicon used in the argumentations. We present the results obtained when evaluating the different methods we developed, together with a discussion on the issues encountered and some possible solutions. We conclude that, even at a more general level, our approach performs better than trained classifiers on specific data. For details, please see [7].

7 Conclusions and Future Work

In this paper we have described our present contribution to the task of opinion mining in texts pertaining to different languages and text genres. We have seen that opinion-related tasks are complex tasks and that there is a need for resources, methods and tools that are working well, both individually as well as in processing pipelines. On the other hand, there are many possibilities to apply the opinion mining techniques to real (web or local) applications that can help individuals, organizations in many domains of interest - economic, social, politic etc..

Generally, future work includes the creation and/or testing of alternative resources and tools for sentiment mining. Thus, we can have a measure of the degree in which each of the components we use contributes to the success or failure of the system. Moreover, the impact of different components and subtasks of the systems must be evaluated (such as anaphora resolution, coreference resolution on the task of opinion classification in different text genres), in order to assess the importance and cost of their use.

References

1. Balahur, A.; Montoyo, A. Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification. In review Procesamiento del Lenguaje Natural ,Vol: 40, Num: 40, 2008.
2. Balahur, A.; Montoyo, A.: An Incremental Multilingual Approach to Forming a Culture Dependent Emotion Triggers Lexical Database. In: Proceedings of the Conference of Terminology and Knowledge Engineering (TKE 2008).
3. Balahur, A.; Montoyo, A. Multilingual Feature-Driven Opinion Extraction and Summarization from Customer Reviews. Lecture Notes in Computer Science , Vol: 5039 , Num: 5039, 2008.
4. Balahur, A.; Montoyo, A. Building a Recommender System Using Community Level Social Filtering. In Proceedings of the 5th International Workshop on Natural Language and Cognitive Science. Barcelona, Spain, 2008.
5. Balahur, A.; Montoyo, A. Determining the Semantic Orientation of Opinions on Products - a Comparative Analysis. In review Procesamiento del Lenguaje Natural ,Vol: 41, Num: 41, 2008.
6. Balahur, A.; Lloret, E.; Ferrández, O.;Montoyo, A., Palomar, M.; Muñoz,R: The DLSIUAES Team's Participation in the TAC 2008 Tracks. In Proceedings of the Text Analysis Conference 2008 Workshop, 17-19 Nov. 2008, Washington, USA.
7. Balahur, A.; Kozareva, Z.; Montoyo, A: Determining the Polarity and Source of Opinions Expressed in Political Debates. In Proceedings of CICLing 2009.
8. Strapparava, C. and Valitutti, A. "WordNet-Affect: an affective extension of WordNet". In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083-1086.
9. Scherer, K. and Wallbott, H.G. The ISEAR Questionnaire and Codebook, 1997.
10. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines. Microsoft Research Technical Report MSRTR- 98-14 (1998)
11. Cilibrasi, D., Vitanyi, P.: Automatic Meaning Discovery Using Google. IEEE Journal of Transactions on Knowledge and Data Engineering (2006)

12. Deerwester, S. Dumais, S., Furnas, G. W., Landauer, T. K, Harshman, R.:Indexing by Latent Semantic Analysis.In: Journal of the American Society for Information Science 41 (6): 391-407.
13. Diccionario Ideológico de la Lengua Española, Larousse Editorial, RBA Promociones Editoriales, S.L., ISBN 84-8016-640
14. Maslow, A.H. 1943. A Theory of Human Motivation. Psychological Review 50 (1943):370-96.
15. Max-Neef, M. A. 1991: Human scale development: conception, application and further reflections. The Apex Press. New York
16. Sperber, D., Wilson, D.2004. Relevance Theory. In G. Ward and L. Horn (eds) Handbook of Pragmatics. Oxford: Blackwell, pp. 607-632.
17. Stevenson, C. 1963. Facts and Values: Studies in Ethical Analysis. Yale University Press, New Haven, USA.

Clasificación Ordinal de Documentos Según Grado de Sentimiento y de Influencia

Emili Sapena, Lluís Padró, Jordi Turmo
TALP Research Center
Universitat Politècnica de Catalunya
Barcelona, Spain
{esapena, padro, turmo}@lsi.upc.edu

No Institute Given

1 Introducción

El trabajo de investigación realizado tiene como objetivo asignar un grado de sentimiento y de influencia a documentos en lenguaje natural no estructurado. Se proponen varias aproximaciones de aprendizaje automático supervisado y distintos modelos atributivos para este tipo de clasificación ordinal.

2 Dimensión 1: Grado de Influir

La dimensión Grado de Influir valora el deseo de influir al lector que tiene un usuario al escribir un texto. Es decir, se valora en qué medida el texto se ha escrito buscando que su opinión cree reacción en quien la lea, ya sea positiva o negativa, y si intenta involucrar a todo el que lo lea para que se posicione. También se tiene en cuenta la posición opuesta, es decir, los textos con preguntas o peticiones sobre temas en los que no se tiene una opinión firme. Éstos últimos se considera que tienen un grado negativo porque desean ser influidos.

Idealmente, la dimensión se define como una línea donde los valores positivos representan la intención de querer influir sobre los demás, los negativos la predisposición a ser influido y los valores cercanos a cero indican neutralidad. Cuanto mayor sea el valor, mayor es el deseo de influir o querer ser influido, en función del signo.

3 Dimensión 2: Grado de Sentimiento

La dimensión Grado de Sentimiento valora la positividad o negatividad en la manera de formular una opinión. Independientemente de que se esté opinando sobre un concepto concreto negativa o positivamente, el sentimiento positivo o negativo que transmite el documento es lo que se valora en esta dimensión. Se tiene en cuenta sobretodo la connotación positiva o negativa de las formas utilizadas ya sean adjetivos, nombres, verbos o adverbios.

Idealmente, la dimensión se define como una línea donde los valores positivos representan los textos con positividad, los negativos representan negatividad y el cero, o los valores cercanos, neutralidad.

4 Corpus

Los documentos que se han utilizado en los experimentos han sido extraídos "tal cual" de foros en Internet de usuarios de ciertos productos. Al no tratarse de textos regulares como pueden ser los libros o los artículos de periódico, éstos contienen errores, faltas de ortografía y gramaticales, palabras no incluidas en diccionarios, jergas locales, tacos y, en general, incorrecciones de todo tipo. Todo ello, por un lado, dificulta la tarea del análisis morfológico y sintáctico así como la detección de entidades. Mientras que por otro lado añade viveza e indicaciones que un texto regular no tendría como son las repeticiones de palabras, exclamaciones, mayúsculas y emoticonos (smyleis).

El corpus utilizado consta de 500 documentos anotados manualmente para ambas dimensiones. Para la dimensión Grado de Influir se han utilizado 3 etiquetas: I1 (deseo de ser influido), I2 (neutralidad), I3 (deseo de influir). Y para la dimensión Grado de Sentimiento se han utilizado 5: S1 y S2 (sentimiento negativo), S3 (neutro), S4 y S5 (sentimiento positivo).

5 Algoritmos Aprendizaje Supervisado

A continuación se explican cada una de las aproximaciones usadas. Se han probado algoritmos genéricos para clasificar en varias clases, y otros que tienen en cuenta que las clases finales están ordenadas.

- **SVM Multiclase**: Se trata de la aproximación más directa. Se entrena un clasificador de N clases para que clasifique cada documento en su clase final correspondiente. Los clasificadores multiclase no aprovechan el hecho de que las clases finales esten ordenadas
- **Binarias**: El modelo binario entrena un clasificador distinto para cada clase. Cada clasificador decide si un documento pertenece a su clase o no devolviendo una probabilidad. Al documento se le asigna la clase con mayor probabilidad. Este modelo tampoco tiene en cuenta que las clases finales estén ordenadas.
- **SVOR: Support Vector Ordinal Regression**. Basado en las dos aproximaciones presentadas por [1]. Se trata de dos kernels que aprovechan el hecho de que las clases finales tienen un orden. La ventaja está en que la condición de que las clases están ordenadas se encuentra en el kernel y no en el modelo. La utilización es la misma que el clasificador multiclase. Los dos kernels son:
  - SVORIM: La condición de que las clases finales tienen orden se cumple implícitamente en el kernel.
  - SVOREX: La condición de que las clases finales tienen orden se especifica explícitamente al generar el kernel.
- **Espacio embedido**: El clasificador de espacio embedido aprovecha que las clases están ordenadas adaptando el modelo atributivo. Las instancias de entrada contienen los atributos del documento a clasificar y también la posible clase que se le va a asignar. Se trata de un solo clasificador binario que indica si un documento se encuentra en una clase superior o en una inferior a la preguntada. Es decir, dado un documento, éste se pasa por el clasificador preguntando cada vez por una clase distinta. Se asigna la clase definitiva utilizando los límites inferior y superior que indica el clasificador [2].

– **Ordinal por umbrales**: El clasificador por umbrales crea un clasificador binario por cada umbral entre dos clases. De esta forma, cada clasificador binario indica si el documento se debe clasificar por encima o por debajo de dicho umbral. La clase final se obtiene combinando todos los valores obtenidos como si se trataran de probabilidades y asignando el documento a la clase mas probable [3].

6 Modelos Atributivos

Los modelos atributivos usados van aadiendo nuevos atributos respecto a sus modelos anteriores. A continuación se describen los modelos atributivos usados:

– **Modelo Inicial**. Atributos basados en elementos sintácticos, morfológicos y formas concretas.
– **Modelo A1**. Se aaden atributos estadísticos: longitud de frases, número de palabras, etc.
– **Modelo A2**. Se aaden listas de adjetivos, nombres y verbos característicos de cada clase. Las listas se obtienen mediante Información Mútua.
– **Modelo A3**. Se aaden listas de bigramas característicos de cada clase.
– **Modelo A4**. Se aaden listas de coocurrencias de adjetivos, nombres y verbos característicos de cada clase.

7 Experimentos

Se han realizado diversos experimentos para comparar la efectividad de las distintas aproximaciones y para encontrar el modelo atributivo que mejor determina la clase de un documento. En cada experimento se comparan los resultados obtenidos por cada una de las aproximaciones. Las Tablas 1, 2, 3 y 4 muestran los resultados obtenidos para los distintos modelos atributivos. La columna “acc” se refiere al porcentaje de acierto, que se calcula haciendo el porcentaje de documentos que estan bien asignados en su clase respecto al número total de documentos. La columna “err” se refiere a la distancia media de error, que se obtiene sumando todas las distancias de error y dividiendo entre el número total de documentos. Teniendo en cuenta que las clases finales están ordenadas, los errores con mayor distancia se pueden considerar mas graves.

Modelo A1	Grado Influir		Grado Sentimiento	
Aproximación	acc	err	acc	err
baseline	49,6	0,504	49,2	0,598
SVM Multiclase	59,6	0,422	51,0	0,584
SVM Binarias	58,0	0,438	49,4	0,604
SVM Ordinal	58,0	0,428	51,4	0,564
SVORIM	62,0	0,388	49,4	0,604
SVOREX	62,2	0,386	48,8	0,600
SVM Embedido	63,0	0,378	49,2	0,598

Table 1. Resultados de los algoritmos usando el Modelo A1.

Modelo A2	Grado Influir		Grado Sentimiento	
Aproximación	acc	err	acc	err
baseline	49,6	0,504	49,2	0,598
SVM Multiclase	67,0	0,350	61,0	0,482
SVM Binarias	64,6	0,366	58,2	0,502
SVM Ordinal	62,4	0,380	56,4	0,514
SVORIM	67,8	0,330	57,0	0,480
SVOREX	66,0	0,350	57,4	0,472
SVM Embedido	65,2	0,356	53,8	0,522

Table 2. Resultados de los algoritmos usando el Modelo A2.

Modelo A3	Grado Influir		Grado Sentimiento	
Aproximación	acc	err	acc	err
baseline	49,6	0,504	49,2	0,598
SVM Multiclase	76,8	0,244	71,2	0,342
SVM Binarias	67,6	0,326	66,4	0,390
SVM Ordinal	67,2	0,330	64,8	0,400
SVORIM	75,2	0,256	67,2	0,358
SVOREX	74,2	0,270	66,6	0,362
SVM Embedido	72,0	0,284	58,6	0,442

Table 3. Resultados de los algoritmos usando el Modelo A3.

Modelo A4	Grado Influir		Grado Sentimiento	
Aproximación	acc	err	acc	err
baseline	49,6	0,504	49,2	0,598
SVM Multiclase	78,4	0,230	72,6	0,318
SVM Binarias	68,0	0,320	65,6	0,408
SVM Ordinal	66,2	0,338	64,6	0,404
SVORIM	76,2	0,242	68,4	0,348
SVOREX	77,6	0,228	67,8	0,352
SVM Embedido	73,6	0,268	59,6	0,428

Table 4. Resultados de los algoritmos usando el Modelo A4.

References

1. Chu, W., Keerthi, S.: New Approaches to Support Vector Ordinal Regression. In: International Conference on Machine. Volume 2005. (2005) 07–11
2. Rajaram, S., Garg, A., Zhou, X., Huang, T.: Classification Approach towards Ranking and Sorting Problems. LECTURE NOTES IN COMPUTER SCIENCE (2003) 301–312
3. Frank, E., Hall, M., of Waikato, U., of Computer Science, D.: A Simple Approach to Ordinal Classification. LECTURE NOTES IN COMPUTER SCIENCE (2001) 145–156

Las Tecnologías del Lenguaje Humano en la comprensión de los diferentes registros del lenguaje*

Ester Boldrini, and Patricio Martínez-Barco

Depto. Lenguajes y Sistemas Informáticos
Universidad de Alicante,
Carretera San Vicente del Raspeig s/n - 03690 San Vicente del Raspeig – Alicante,
{eboldrini, patricio}@dlsi.ua.es

Resumen. El objetivo de este artículo es presentar los planes de investigación en el ámbito del estudio, comprensión y formalización de las emociones en distintos registros textuales. Se mencionará brevemente el estado de la cuestión y asimismo se ilustrarán los planes de investigación a largo término y las investigaciones a corto plazo que están estamos desarrollando, haciendo hincapié en las posibles dificultades y proponiendo nuestras soluciones para resolver dichos obstáculos.

Palabras clave: Emociones, información sujetiva, múltiples registros de uso, multilingualidad, multimodalidad, blogs, foros, modelo de anotación.

1 Introducción

Según nuestro conocimiento, en la actualidad no existe un modelo global multilingüe para la representación de las opiniones y de las emociones que tenga una aplicación directa en la comprensión automática del lenguaje. Nuestro objetivo principal es crear un modelo exhaustivo que sea capaz de entender las emociones tratadas en distintos géneros textuales y en diferentes lenguas. Dicho modelo se aplicará a sistemas de aprendizaje automático (AA) y se hará que la máquina aprenda los conceptos o rasgos que hemos anotado previamente y así poderlos detectar en otros corpus planos.

* Esta investigación ha sido financiada por los proyectos TEXT-MESS (TIN2006-15265-C06-01), por el proyecto QALL-ME (FP6 IST-033860) y por la beca de iniciación a la investigación del Vicerrectorado de investigaci´n de la Universidad de Alicante (ref BII2008-7898717).

Los géneros textuales que analizaremos no serán los convencionales, sino que centraremos nuestra atención en los géneros de la Web, como blogs, foros, chats, etc.
Cabe destacar que la tarea de comprensión de la información no sólo se centra en la comunicación formal explícita, sino también en la información más allá del texto, es decir la situación pragmática.
Hemos optado por géneros textuales no convencionales, dado que la información colaborativa, está empezando a ser utilizada como referencia por organizaciones y particulares, y por lo tanto es la protagonista de las páginas de Internet.
El artículo está organizado como sigue: en la sección 2 se presenta un breve estado de la cuestión y en la sección 3 se ilustra nuestra propuesta. A continuación, en el apartado 4 describimos nuestra metodología de investigación y en la sección 5 presentamos el trabajo que estamos desarrollando.

2 Estado de la cuestión

Recientemente, el estudio de las emociones, su comprensión y formalización ha sido un tema de gran interés en la investigación mundial.
Entre otros trabajos, podemos mencionar a Kushal Dave, Steve Lawrence y David Pennock [1] que realizan un estudio sobre extracción de clasificación semántica de opiniones sobre productos.
En [2], H. Yu y V. Hatzivassiloglou nos presentan un trabajo en el cual las opiniones se separan de los hechos y se determina la polaridad en las frases que expresan dichas opiniones.
Asimismo, Wilson, Wiebe y Hwa [3] clasifican el grado de opiniones y otros tipos de elementos sujetivos, mientras que Stoyanov, Cardie, Wiebe y Litman [4] evalúan un esquema de anotación utilizando un corpus de preguntas y respuestas.
En [5], Bethard, Yu, Thornton, Hativassiloglou, y Jurafsky crean un sistema automático de extracción de opiniones y en [6] Wiebe y Mihalcea demuestran que la subjetividad es una propiedad que puede estar relacionada con el sentido de las palabras y la desambiguación de sentidos se puede beneficiar de la anotación de la subjetividad.
Cabe destacar la investigación de Wiebe, Wilson y Cardie [7], que crean un corpus recopilando artículos de prensa para la anotación de las emociones y en otro trabajo Rloff, y Wiebe [8] presentan un sistema de *Information Extraction* (IE) que utiliza un clasificador de frases sujetivas para filtrar sus extracciones.
Entre las investigaciones más recientes podemos apreciar [9] en el que los autores proponen una nueva manera de acercarse al análisis de los sentimientos. Gracias al sistema que crean, es posible identificar automáticamente la polaridad para una amplia variedad de expresiones sujetivas, obteniendo buenos resultados.
Finalmente, Somasundaran, Ruppenhofer y Wiebe [10] están convencidos de que la determinación automática de los roles semánticos no es suficiente. Se basan sobre la experiencia de anotación manual de opiniones y presentan un estudio sobre la atribución de opiniones a fuentes.

Para concluir esta sección, cabe destacar *Swotti*[†], un interesante servicio creado por la española BuzzTrend[‡]. Este proyecto nace con el objetivo de ser un buscador que rastree toda la información contenida en la red. Toma en consideración sólo informaciones que recojan la opinión de los usuarios. En otras palabras se podría definir como un buscador de opiniones.
Esta herramienta trabaja con la información de la Web semántica, una tecnología que puede identificar los adjetivos y verbos que definen el objeto que estamos buscando, y que por tanto permiten deducir si el comentario es positivo o negativo. Haciendo una búsqueda en *Swotti* obtenemos no sólo resultados, sino sobre todo una valoración cualitativa del producto que estamos buscando. Todo esto puede aplicarse a personas, marcas, productos, empresas, ciudades, etc.

3 Propuesta de investigación

Nuestra investigación está enmarcada en el estudio de las técnicas del Procesamiento del Lenguaje Natural (PLN) aplicado a la detección y formalización de todas aquellas estrategias lingüísticas que los hombres utilizamos para expresar nuestras preferencias, emociones o en general enunciados sujetivos que tengan distintos rasgos, dependiendo del género y de la temática que se está tratando.
Más concretamente, nuestro objetivo es analizar distintas tipologías textuales disponibles en Internet, así como diálogos multilingües. Asimismo, para que el estudio sea completo y lo más eficaz posible se hará especial hincapié en el análisis pragmático. En efecto, estamos convencidos de que el contexto discursivo puede ser una de las llaves más importantes para una comprensión exhaustiva del discurso y para una correcta interpretación de ello y como consecuencia poder llegar a formalizarlo.
Como podemos deducir, el presente trabajo de investigación tiene la característica fundamental de multidisciplinariedad; se ven implicadas distintas áreas de investigación, como por ejemplo la Lingüística, la Inteligencia Artificial, la Traducción y también la Psicología. Por lo tanto, nos encontramos ante la necesidad de enfocar el trabajo de una manera interdisciplinaria y asimismo estamos convencidos de que dicha característica representa una ventaja para poder llegar a crear un modelo cualitativa y cuantitativamente representativo gracias al que se pueda detectar la formulación de opiniones sujetivas.
Una de nuestras principales metas consiste en extraer características de los distintos géneros textuales que nos permitan distinguir la narración objetiva de los enunciados sujetivos y, una vez alcanzada esta meta, llegar a un nivel más profundo que nos permita clasificar el tipo de información sujetiva que se extrae del texto, determinar su polaridad y clasificar sus rasgos distintivos, todo esto gracias al utilizo de un modelo de anotación que crearemos *ad hoc* para la detección de dichos rasgos.
Asimismo, nos gustaría destacar la complejidad del trabajo que será llevado a cabo de una manera multilingüe con especial interés en Español, Italiano e Inglés y asimismo

[†]http://www.swotti.com/
[‡] http://www.buzztrend.com/

la complejidad de algunos de los géneros textuales que se tratarán, como por ejemplo blogs o foros en los que son muchas las ocasiones en las que nos podemos encontrar ante numerosas variantes dialectales del mismo idioma. Otro elemento objeto de gran interés en nuestro estudio es el fenómeno lingüístico de la correferencia, en la generación de opiniones y emociones sobretodo a nivel intertextual.

4 Metodología

Por lo que se refiere en general a la metodología que nos hemos planteado seguir, podemos decir que unos de los pasos más importantes consistirá en la recopilación de numerosos corpus de distintas tipologías textuales: artículos de periódicos, de opinión, así como blogs, foros, chats, transcripciones de diálogos, etc. Y además de distintas temáticas que actualmente componen una gran parte de los contenidos existentes en la Web, que es actualmente la mayor fuente de información digitalizada.
No hay duda de que el contenido de la Web ha cambiado; con la *Web 1.0* el contenido mayoritario era básicamente estético y formal, pero gracias a la *Web 2.0* actualmente se encuentran contenidos mucho más dinámicos e informales expresados a través de nuevos registros, por lo que es necesario invertir un esfuerzo adicional en adaptar las técnicas preexistentes hacia estos nuevos registros.
En el paso siguiente se llevará a cabo un análisis profundo de los textos en las distintas lenguas, evidenciando las problemáticas relativas a la detección y clasificación de las sentencias sujetivas, su polaridad, el grado de dicha polaridad y se ofrecerá una propuesta más allá de la simple detección de si una opinión o una expresión es positiva o negativa; se determinará la manera para poder clasificar las opiniones a través de un modelo aplicable a otros corpus.
Como hemos dicho anteriormente, sistemas de *machine learning* (ML) utilizarán la anotación con el objetivo de aprender el modelo del lenguaje; sucesivamente, los resultados obtenidos se podrán aplicar a tareas prácticas como la previsión de las preferencias de los posibles clientes sobre un determinado producto, el sentimiento de los ciudadanos ante la crisis económica o cualquier otra aplicación que tenga ventajas comerciales y de otra naturaleza.
Debemos tener en cuenta una problemática añadida y es que los géneros objeto de nuestro análisis presentan distintos rasgos, aspecto que complica y dificulta la labor investigadora, pero nuestra idea es crear un modelo que no sea demasiado específico, sino único y al mismo tiempo exhaustivo para analizar la mayoría de los géneros textuales disponibles en la Web.
De esta forma podemos definir la creación de un modelo para la comprensión de las emociones a un nivel multilingüe como un proyecto en el que un análisis lingüístico y textual profundo resulta imprescindible para la creación de un modelo efectivo y útil para el ML.

4.1 Objetivos concretos de la investigación

El primer paso de la investigación consiste en realizar un estudio exhaustivo del estado de la cuestión de la materia. Se analizarán los modelos creados para la anotación de opiniones y emociones y además se estudiarán de los esquemas de anotación anafórica existentes y de las distintas situaciones comunicativas.
Después de esta fase preliminar se procederá a la recopilación de corpus paralelos multilingües con las siguientes características: coherencia, adecuación, equilibrio, representatividad. Se detectarán los rasgos comunes entre los géneros y entre las lenguas.
Además, se llevará a cabo un análisis del vocabulario y de los modismos detectados junto con un estudio de los distintos contextos discursivos.
Objetos de nuestro análisis serán también los elementos de correferencia, y se hará hincapié en el estudio de su función a lo largo del texto y sobre todo entre textos.
El paso siguiente consistirá en la creación de un esquema de anotación completo, independiente de la lengua y al mismo tiempo simple de utilizar. Su creación será el resultado de un análisis profundo de los textos a anotar, para que pueda ser lo más completo y adecuado posible a nuestras exigencias y que posteriormente pueda ser evaluado. Además, se procederá a probar este modelo en los sistemas de extracción de opiniones que se están desarrollando actualmente en la Universidad de Alicante.
El objetivo que nos planteamos más a largo plazo consiste en entrenar sistemas de ML para la extracción de la información sujetiva y por lo tanto permitir razonamientos sobre la subjetividad. Todo esto se llevará a cabo evidenciando las diferencias entre los llamados "textos formales" y los nuevos géneros nacidos con la Web.

5 Estado actual de la investigación

Por lo que se refiere a la investigación actualmente en curso, después de haber realizado un estudio exhaustivo del estado de la cuestión, hemos recopilado un corpus trilingüe.
Está compuesto por tres grandes temáticas que son: el protocolo de Kyoto, las elecciones de Zimbabwe y las elecciones de EEUU. Tenemos 30.000 palabras respectivamente para el Español, Italiano e Inglés para cada uno de los temas arriba mencionados que además son de género textual de blog.
Después de la recopilación, hemos realizado el estudio de dichos corpus para tener una idea de los rasgos principales de la expresión de las emociones en esta tipología textual y de allí hemos generado un modelo de anotación.
Gracias a este modelo, tenemos la posibilidad de detectar:

Tabla 1. Breve descripción del modelo: elementos y atributos de cada elemento.

Elementos	Descripción
Discurso objetivo	fuente
Discurso sujetivo	grado, emoción, fenómeno, polaridad, y fuente

Adjetivos	grado, emoción, fenómeno, modificador o no polaridad, y fuente
Adverbios	grado, emoción, fenómeno, modificador o no polaridad, y fuente
Preposiciones	grado, emoción, fenómeno, modificador o no polaridad, y fuente
Verbos	grado, emoción, fenómeno, polaridad, y fuente
Anáfora	tipo y fuente
Pal. en mayúscula	grado, emoción, fenómeno, modificador o no polaridad, y fuente
Puntuación	grado, emoción, fenómeno, modificador o no polaridad, y fuente
Extranjerismos	Latín o inglés grado, emoción, fenómeno, modificador o no polaridad, y fuente
Nombres	grado, emoción, fenómeno, modificador o no polaridad, y fuente
Errores ortográficos	Corrección, grado, emoción, fenómeno, modificador o no polaridad, y fuente
Tipo de frase	Frase, coloquialsimo, modismo, vulgarismo, título
Emociones	Preference, sarcasm, irony, scepticism, accept, anger, bad, confidence, correct, criticism, disgust, excuse, fear, force, good, important, incorrect, interesting, joy, justice, objection, opposition, purity, sadness, support, surprise, guilt, shame, thank, trust, unimportant, will, yield, joke, revendication, envy, rivalry, jealousy, compassion, anxiety, mourning, troubledness, grief, lament, depression, vexation, despondency, sluggishness, fright, timidity, consternation, bewilderment, revenge, rage, hatred, enmity, wrath, greed, longing, malice, rapture, smug, ostentation, anticipation, disappointment, remorse, respect, patience, appreciation, hope, warning, confort,

disconfort, and rejection.

Bibliografía

1. Kushal Dave, Steve Lawrence, and David Pennock. 2003. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. In International World Wide Web Conference, pages 519–528.
2. Yu H. and Hatzivassiloglou V. (2003): *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*. In Proceedings of EMNLP.
3. Wilson, T. and Wiebe, J. (2004): *Just how made are you? Finding strong and weak opinion clauses*. Proc. 19th National Conference on Artificial Intelligence (AAAI-2004).
4. Stoyanov V., Cardie C., Litman D., and Wiebe J. (2004): *Evaluating an Opinion Annotation Scheme Using a New Multi-Perspective Question and Answer Corpus*. AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications.
5. Bethard S., Yu H., Thornton A., Hativassiloglou V., and Jurafsky D. (2004): *Automatic extraction of opinion propositions and their holders*. In 2004 AAAI Spring Symposium on ExploringAttitude and Affect in Text.
6. Wiebe, J. and Mihalcea R. (2006): *Word Sense and Subjectivity*. ACL-2006.
7. Wiebe J., Wilson T. and Claire Cardie (2005): *Annotating expressions of opinions and emotions in language*. Language Resources and Evaluation, 1(2).
8. Riloff E. and Wiebe J. (2003): *Learning extraction patterns for subjective expressions*. In Proceesings of EMNLP.
9. Wilson T., Wiebe J. and Hoffmann P. (2008): *Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis*. To appear in Computational Linguistics.
10. Somasundaran S., Wiebe and Ruppenhofer J. (2008): *Discourse Level Opinion Interpretation*. COLING-2008.

Monolingual and Crosslingual Plagiarism Detection

Towards the Competition @ SEPLN09 *

Alberto Barrón-Cedeño and Paolo Rosso

Natural Language Engineering Lab, RFIA,
Departmento de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia
{*lbarron, prosso*}*@dsic.upv.es*

Abstract. Automatic plagiarism detection considering a reference corpus compares a suspicious text to a set of documents in order to relate the plagiarised fragments to their potential source. The suspicious and source documents can be written wether in the same language (monolingual) or in different languages (crosslingual).
In the context of the Ph. D., our work has been focused on both monolingual and crosslingual plagiarism detection. The monolingual approach is based on a search space reduction process followed by an exhaustive word n-grams comparison. Surprisingly it seems that the application of the reduction process has not been explored in this task previously. The crosslingual one is based on the well known IBM-1 alignment model. Having a competition on these topics will make our work available to the Spanish scientific community interested in plagiarism detection.

1 Introduction

The easy access to a wide range of information in multiple languages via electronic resources has favoured the increase of text plagiarism cases of both kinds: monolingual and crosslingual. To plagiarise means to use text written by other people (even adapting it by rewording, insertion or deletion) without credit or citation. From a crosslingual perspective, a text fragment in one language is considered a plagiarism of a text in another language if their contents are considered semantically similar no matter they are written in different languages.

In order to get enough evidence to prove if a text is plagiarised, it is necessary to find its potential source. The objective of plagiarism detection with reference is to give this evidence. This is carried out by searching for the potential source of a suspicious text fragment from a set of reference texts.

Few works have been made from a crosslingual point of view. The first one is based on explicit semantic analysis, where two comparable corpora (one on each implied language) are exploited in order to define how semantically closed two

* We would like to thank the MCyT TEXT-MESS CICYT TIN2006-15265-C06-04 research project as well as the TIMM network CICYT TIN2005-25825-E.

documents are [11]. The second one is based on statistical bilingual models [4, 9] (Section 3). Note that no translation process is carried out in both approaches.

With the aim of bringing together to the researchers interested in these topics, we plan to carry out a competition which will be held in the context of the proposed PAN Satellite Workshop of the SEPLN'09 conference.

2 Monolingual Plagiarism Detection

An important factor in the plagiarism detection with reference is precisely the reference corpus. The best available method would be useless if the source of a plagiarised text is not included into the reference corpus D. Due to this reason, reference corpora are composed of a huge set of potential source documents.

Comparing a suspicious text s to all the reference documents $d \in D$ is practically impossible. Our proposed method carries out a preliminary reduction process, based on the Kulback-Leibler distance, selecting only those documents d with a high probability of being the source of s [3, 1]. Each probability distribution P_d is compared to the probability distribution P_s. The ten most similar reference documents are considered as candidates of being the source of the potentially plagiarised sentences in s. This is the reduced reference set D'.

The following objective is to answer the question "*Is a sentence $s_i \in s$ plagiarised from a document $d \in D'$?*". Due to the fact that plagiarised text fragments use to be rewritten from their source, a rigid search strategy does not give good results. Our flexible search strategy is based on a word n-grams comparison [2]. We consider n-grams due to the fact that independent texts have a small amount of common word n-grams (considering $n \geq 2$).

Our approach is based on the comparison of suspicious sentences and reference documents. We do not split the reference documents into sentences due to the fact that a plagiarised sentence could be made of fragments from multiple parts of a source document. The basic schema is as following: (1) s is split into sentences (s_i); (2) s_i is split into word n-grams, resulting in the set $N(s_i)$; (3) $d \in D'$ is not split into sentences, but simply into word n-grams, resulting in the set $N(d)$; and (4) $N(s_i)$ is compared to $N(d)$. Due to the difference in the size of $N(s_i)$ and $N(d)$, an asymmetric comparison is carried out on the basis of the *containment* measure [7]:

$$C(s_i \mid d) = \frac{|N(s_i) \cap N(d)|}{|N(s_i)|} \tag{1}$$

If the maximum $C(s_i \mid d)$, after considering every $d \in D'$, is greater than a given threshold, s_i becomes a candidate of being plagiarised from d.

For our experiments we have used the *METER corpus* [6]. This corpus is not a real plagiarism corpus. It is composed of a set of journalistic notes and was originally created in order to analyse the reuse of information in the British newspapers. The interesting fact about this corpus is that the text of a set of newspaper (suspicious) notes is identified as *verbatim*, *rewrite* or *new*, for exact copy, rewritten or nothing to do with the Press Association (reference) notes.

Our experiments show that the best results for the exhaustive comparison are obtained by considering bigrams and trigrams. In both cases, the word n-grams are short enough to handle modifications in the plagiarised sentences and long enough to compose strings with a low probability of appearing in any (but the plagiarism source) text. Trigram based search is more rigid, resulting in a better Precision. Bigram based search is more flexible, allowing better Recall. The search space reduction process improves the obtained F-measure (from 0.68 to 0.75 for bigrams) and the time it takes to analyse a suspicious document is reduced (from 2.32 to only 0.19 seconds in average).

3 Crosslingual Plagiarism Detection

Given the suspicious and reference texts x and y (written in different languages), the objective is to answer the question "*Is x plagiarised (and translated) from y?*". In some way, crosslingual plagiarism analysis is related to crosslingual information retrieval [10]. In fact, the aim is to retrieve those fragments that have been plagiarised in a language with respect to the one originally employed.

In our current research [4, 9] we have composed a minicorpus of original-plagiarised text pairs. The original fragments (y), in English, were extracted from a set of documents on Information Retrieval written by one only author. Around ten plagiarised versions of each fragment y have been obtained in Spanish and Italian (x). Each fragment x has been created by a different "human plagiariser" or automatic machine translator.

The set of y-x pairs was divided into training and test subsets. The training subset was used in order to compose a statistical bilingual dictionary. This dictionary was created on the basis of the IBM-1 alignment model [5], commonly used in statistical machine translation. The test set was only composed of the suspicious fragments from the test pairs. In order to obtain a realistic experiment, text fragments originally written in Spanish (and Italian) were added.

The objective of our experiment was to know if a suspicious fragment x was a plagiarism case from one of our reference fragments y. In order to determine if x is plagiarised from any y fragment, we compute the probability $p(y \mid x)$ of each fragment y given x. This probability is calculated as in Eq. 2.

$$p(y \mid x) = \frac{1}{(|x|+1)^{|y|}} \prod_{i=1}^{|y|} \sum_{j=0}^{|x|} p(y_i \mid x_j) \tag{2}$$

where $p(y_i \mid x_j)$ is simply calculated on the basis of the statistical bilingual dictionary previously obtained and $|\cdot|$ is the length of $\cdot$ in words.

Our proposal calculates the probabilistic association between two terms in two different languages. After considering this probability, we are able to determine how likely is that a fragment x is a translation (plagiarism) from y. If the maximum $p(y \mid x)$ after considering every reference fragment y is higher than a given threshold, we consider that x is plagiarised from y.

The results obtained up to now with this method are promising. The application of a statistical machine translation technique, has demonstrated to be a valuable resource for the crosslingual plagiarism analysis. Due to the fact that we determine the similarity between suspicious and original text fragments on the basis of a dictionary, the word order is not relevant and we are able to find good candidates even when the plagiarised text has been modified.

4 Current and Future Work

Currently, we are creating corpora containing both kinds of plagiarism cases. These corpora will be used during the proposed competition as well as for our own research work. We plan also to tackle the problem of plagiarism of ideas "...in which an original thought from another is used but without any dependence on the words or form of the source..." [8]. This is a more general (and hard to detect) case of plagiarism.

References

1. Barrón-Cedeño, A. 2008. Detección automática de plagio en texto. Master's thesis, Universidad Politécnica de Valencia.
2. Barrón-Cedeño, A., Rosso, P.: On Automatic Plagiarism Detection based on n-grams Comparison. In: ECIR. LNCS, *in press* (2009)
3. Barrón-Cedeño, A., Rosso, P., Benedí, J.M.: Reducing the Plagiarism Detection Search Space on the basis of the Kullback-Leibler Distance. IN: CICLing 2008. LNCS, *in press* (2009)
4. Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A.: On Crosslingual Plagiarism Analysis Using a Statistical Model. In: ECAI'08 PAN Workshop Uncovering Plagiarism, Authorship and Social Software Misuse, pp. 9–13, Patras, Greece (2008)
5. Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., Roossin, P.: A Statistical Approach to Machine Translation. Computational Linguistics, 16(2), 79-85 (1990)
6. Clough, P., Gaizauskas, R., Piao, S.: Building and Annotating a Corpus for the Study of Journalistic Text Reuse. In: 3rd International Conference on Language Resources and Evaluation (LREC-02), vol. V, pp. 1678–1691. Las Palmas, Spain (2002)
7. Lyon, C., Malcolm, J., Dickerson, B.: Detecting Short Passages of Similar Text in Large Document Collections. In: Conference on Empirical Methods in Natural Language Processing, pp. 118–125. Pennsylvania (2001)
8. Martin, B.: Plagiarism: a Misplaced Emphasis. Information Ethics. 3(2) 36-47 (1994)
9. Pinto, D., Civera, J., Juan, A., Rosso, P., Barrón-Cedeño, A.: A Statistical Approach to Crosslingual Natural Language Tasks. In: Fourth Latin American Workshop on Non-Monotonic Reasoning, Puebla, Mexico (2008)
10. Pinto, D., Juan, A., Rosso, P.: Using Query-Relevant Documents Pairs for Cross-Lingual Information Retrieval. In: TSD 2007. LNAI 4629, pp. 630–637 (2007)
11. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-Based Multilingual Retrieval Model. In: ECIR 2008. LNCS, vol. 4956, pp. 522–530. Springer-Verlag (2008)

Proyecto de Tesis: Tecnologías del lenguaje humano aplicadas a la atribución de autoría*

M. Pardiño, A. Suárez y P. Martínez-Barco

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
España
{maria,armando,patricio}@dlsi.ua.es

Resumen En este trabajo se presenta una comparativa de las técnicas basadas en compresión y de aprendizaje automático aplicadas en la atribución de autoría. Para ello, hemos realizado una serie de experimentos con distintos algoritmos utilizando diferentes marcas idiosincráticas para analizar la estabilidad tanto de los métodos como de dichas marcas. Finalmente, los resultados obtenidos no permiten concluir que una tipología de técnicas de clasificación sea más efectiva que otra en nuestra evaluación, pero sí han mostrado la mayor estabilidad y escalabilidad de determinados algoritmos y características.

1. Introducción y estado de la cuestión

La atribución de autoría (AA) trata de obtener una "huella"lingüística o perfil de un autor utilizando para ello marcas idiosincráticas que no estén bajo su control, de esta forma, el objetivo final consiste en clasificar documentos por autor. Se trata de un área multidisciplinar en la que se entrecruzan otras áreas de investigación (lingüística, derecho e informática) que trabajan conjuntamente para dotar de herramientas y metodologías apropiadas que permitan automatizar el tratamiento lingüístico en ámbitos jurídicos y judiciales.
Uno de los mayores problemas que presenta la AA es que no siempre es sencilla la reconstrucción del perfil lingüístico del autor, puesto que es posible que éste cambie en función del género o tema, de la época en que fue escrito o incluso dentro del mismo documento según la sección. En ocasiones, también se puede apreciar que una determinada obra ha sido escrita por más de un autor. Otro inconveniente que dificulta los avances en este área es la escasez de corpus estándar que permitan valorar las mejoras introducidas y compararlas con las técnicas existentes.
Cabe mencionar que los primeros estudios para abordar esta tarea estaban basados en la aplicación de técnicas estadísticas [10,4]. Más recientemente se han

* Este trabajo ha sido parcialmente financiado por el gobierno español a través del proyecto TIN-2006-15265-C06-01, del proyecto GV06028, el proyecto QALL-ME, perteneciente al 6º Programa Marco de la Unión Europea (EU), número de contrato: FP6-IST-033860, y el gobierno español con la beca de investigación AP2007-03072 del Programa de FPU del Ministerio de Ciencia e Innovación.

introducido técnicas de aprendizaje automático como redes neuronales [6], clasificadores bayesianos [2], support vector machines [3] y árboles de decisión [8]. En los últimos años han comenzado a aplicarse, también, técnicas de compresión, no estando éstas exentas de controversia [9,1,5,7], dando cabida a nuevos experimentos con el objeto de clarificar los resultados poco concluyentes obtenidos hasta el momento. Tanto los métodos estadísticos como los basados en aprendizaje automático utilizan una serie de marcas para caracterizar el estilo de escritura del autor, mientras que los algoritmos de compresión se aplican sobre los documentos. En este sentido, se utilizan una gran variedad de marcas a distintos niveles lingüísticos (a nivel de token, sintacticos, basados en la riqueza del vocabulario, según la frecuencia de aparición de las palabras, errores ortográficos y gramaticales, etc).
Este trabajo se centra en presentar una primera aproximación realizada dentro del proyecto de tesis del trabajo recién iniciado, que aborda el estudio de técnicas de procesamiento de lenguaje natural (PLN) aplicadas a la AA. A continuación veremos la propuesta inicial, los resultados obtenidos y las líneas futuras de investigación.

2. Comparación de métodos para la Atribución de Autoría

En este trabajo se ha realizado una comparativa de distintas técnicas basadas en Aprendizaje Automático (ML) y técnicas de compresión con el objetivo de encontrar aquellas más adecuadas en la atribución de autor.
Junto con las técnicas basadas en ML se utilizan marcas de estilo para caracterizar la escritura del autor. Estas marcas han sido seleccionadas en la búsqueda de aquellas más características del estilo, así como con el objetivo de analizar la estabilidad de los métodos de ML. En relación a las marcas utilizadas es interesante comentar que son independientes de la lengua, y por tanto, no son necesarios recursos específicos para su adaptación a otros idiomas.
Por un lado, se ha realizado una selección de marcas estilométricas (todas las palabras; las M palabras más frecuentes con M=500, M=5000; las P palabras con mayor información mutua con P=500, P=5000; N-gramas de palabras con N=1, 2 y 3; los M N-gramas más frecuentes con N=1, 2 y 3, M=500, M=5000 y function words) que de forma individualizada han sido combinadas con dos algoritmos de aprendizaje automático, Naïve Bayes y SVM. Por otra parte, se han aplicado tres métodos basados en compresión (rar, gzip y ppmd) directamente sobre los textos. De esta forma, se pretende mostrar una comparativa de distintas metodologías para la caracterización de autor.
Para la validación de los resultados obtenidos, se ha utilizado el corpus Gutenberg634 [11] confeccionado por Zhao utilizando 634 libros en inglés de varios autores incluidos en el proyecto Gutenberg. De esta colección hemos seleccionado 3 subconjuntos de 20, 10 y 5 autores con 10 textos por autor para la realización de nuestros experimentos. Se analizan las distintas propuestas de clasificación con el método de evaluación k-cross fold-validation (con K=10).

3. Experimentos y resultados

En relación al porcentaje de acierto en función del número de autores, cabe destacar que las diferencias más importantes entre los distintos métodos son más fácilmente observables a mayor cantidad de autores, mientras que los resultados se igualan en gran medida para las distintas metodologías con un número pequeño de autores entre los que discernir.

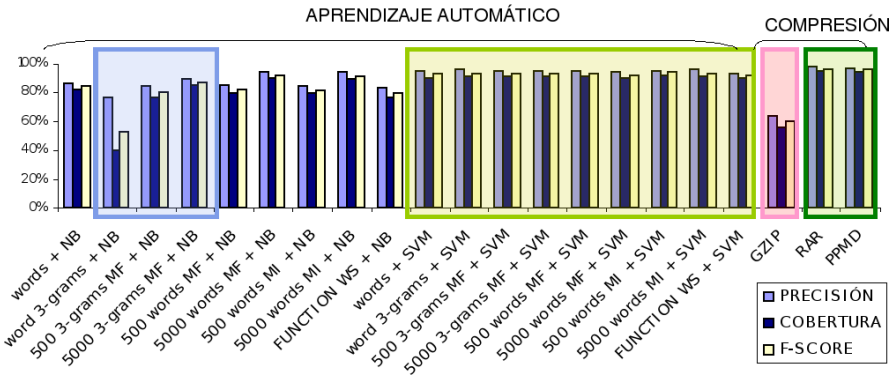

Figura 1. Comparativa de los resultados obtenidos calculados según macropromedio para los distintos métodos propuestos con 20 autores.

En concreto, con un número elevado de autores, SVM obtiene mejores resultados, mientras que con 5 autores está muy igualado con Naïve Bayes. Por otra parte, gzip ha sido una de las técnicas con resultados más pobres (55-64 % de acierto con 20 autores, 64-71 % con 10 autores, y 82-86 % con 5 autores). Al contrario de lo que sucede con gzip, los algoritmos de compresión rar y ppmd obtienen los mejores resultados.
En relación a las marcas de estilo utilizadas, parece que emplear todos los N-gramas de palabras (N=1, 2 y 3) de los textos empeora los resultados e incrementa en gran medida la complejidad de cálculo. Finalmente, no podemos concluir que un tipo de clasificación (basados en aprendizaje automático o en compresión) sea mejor que otro, pero sí podemos afirmar que para el corpus utilizado, diferentes técnicas han obtenido resultados más destacados (en concreto, SVM, rar y ppmd).

4. Conclusiones y trabajo futuro

A pesar de ser un área objeto de estudio durante décadas, la AA está realmente en la fase inicial de su desarrollo. De hecho, no existe un acuerdo claro entre las medidas estilísticas más adecuadas ni entre la metodología de clasificación a aplicar en la tarea.

Uno de lo objetivos que nos planteamos dentro del proyecto de tesis es la búsqueda de las marcas de estilo que mejor representen a un conjunto de autores. De esta forma, nuestro objetivo es analizar características a niveles lingüísticos de mayor complejidad para comprobar su aplicabilidad en AA. Aspectos que requerirán un estudio pormenorizado son la recopilación de nuevos corpus de aprendizaje y evaluación, así como la aplicación de estas técnicas en escenarios multilingües.

Por último, en relación a los buenos resultados obtenidos, éstos se han debido en gran medida, a la utilización de textos de gran tamaño, lo que ha facilitado la clasificación de los mismos. Tras comprobar la eficacia y eficiencia de los distintos métodos sobre un corpus con textos de prueba y entrenamiento de tamaño considerable, consideramos interesante ver el comportamiento de estas mismas técnicas sobre textos más reducidos, puesto que son los casos que se suelen presentar en las situaciones reales de aplicación de la AA en el mundo de la lingüística forense, como son casos de detección de autor en notas de suicidio, cartas de extorsión, etc.

Referencias

1. D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Physical Review Letters*, 88(4):048702, Jan 2002.
2. R.M. Coyotl-Morales, L. Villaseñor-Pineda, M. Montes-y Gómez, and P. Rosso. Authorship attribution using word sequences. In *Proceedings of the 11th CIARP Iberoamerican Congress on Pattern Recognition*, pages 844–853. Springer, 2006.
3. J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1):109–123, 2003.
4. J.M. Farringdon. Analysing for authorship: A guide to the cusum technique, with contributions of a.q. morton, m.g. farringdon and m.d. baker. *Science and Justice*, 1996.
5. J. Goodman. Extended comment on language trees and zipping. *Condensed Matter Archive, Feb*, 21:0202383, 2002.
6. J.F. Hoorn, S.L. Frank, W. Kowalczyk, and F. van der Ham. Neural network identification of poets using letter sequences. *Literary Linguist Computing*, 14(3):311–338, 1999.
7. D.V. Khmelev and W.J. Teahan. Comment 'language trees and zipping'. *Phys. Rev. Lett.*, 90(2):089803, 2003.
8. M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, pages 69–72, Acapulco, Mexico, 2003.
9. O.V. Kukushkina, A.A. Polikarpov, and D.V. Khmelev. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2):172–184, 2001.
10. M.W.A. Smith. Recent experience and new developments of methods for the determination of authorship. *Association for Literary and Linguistic Computing Bulletin*, 11(3):73–82, 1983.
11. Y. Zhao. *Effective Authorship Attribution in large document collections.* PhD thesis, Computer Science and Information Technology, 2008.

LOQEVAL: Propuesta de evaluación de la calidad de objetos de aprendizajes mediante ontologías

Dianelys Medina*, Jesús M. Hermida** and Andrés Montoyo

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{dianelys, jesusmhc, montoyo}@dlsi.ua.es

Resumen Un aspecto importante de los objectos de aprendizaje en el campo del e-learning, dada la relevancia adquirida en la actualidad, es la evaluación de su calidad. Varios métodos se han desarrollado durante los últimos años que permiten valorarla desde diferentes aspectos. Sin embargo, no existe ninguna aproximación que permita que este proceso se realice de forma automátizada. Es por ello que en este trabajo se presenta LOQEVAL, una propuesta de evaluación de la calidad de objetos de aprendizaje mediante el uso de ontologías, con el fin de cubrir parte de las carencias que se encuentran en los trabajos actuales.

1. Introducción

En el campo del e-learning, como parte de la interoperabilidad entre diferentes plataformas, se busca que los contenidos sean portables, reutilizables e intercambiables entre aplicaciones, esto ha dado origen a los llamados objetos de aprendizajes (OA) [1]. Éstos son unidades de contenido con significado propio, constituidos por paquetes de información multiformato, orientados al logro de un determinado objetivo pedagógicos, identificables por metadatos e integrados por recursos, actividades y evaluación [2]. Teniendo en cuenta la importancia de los OA en el campo del e-learning y su repercusión en la formación de estudiantes, el concepto de calidad en los OA se convierte en un aspecto importante en esta área. Esto conlleva el desarrollo de los mecanismos oportunos para la evaluación de los OA con el objetivo de asegurar la obtención de recursos de aprendizajes con un nivel de calidad determinado.Es por ello que en este trabajo nos centramos en investigar y definir una nueva propuesta de evaluación de calidad de OA mediante ontologías, llamada LOQEVAL.

* Este trabajo ha sido parcialmente financiado por la beca MAEC-AECID, programa II-A.
** Este trabajo ha sido parcialmente financiado por el programa de FPU del Ministerio de Ciencia e Innovación, a través de la beca AP2007-03076.

2. Estado del Arte

Hasta la fecha se han realizado estudios y propuestas para la evaluación de la calidad de los OA, tales como LORI(Learning object review instrument),el cual propone un marco de evaluación de OA basado en el análisis de nueve dimensiones [3]. Esta evaluación individual puede ser perfeccionada con una evaluación colaborativa denominada Modelo de Participación Convergente [4]. Otra propuesta es la desarrollada en [5], en la cual propone criterios para evaluar los objetos de aprendizajes, agrupados en 4 dimensiones o aspectos, donde los evaluadores sugieren un marco de evaluación integral de los OA desde la perspectiva pedagógica, curricular, técnica y funcional. En [6] se centra el estudio de la calidad en los OA como producto (el objeto mismo) o como proceso (desarrollo del objeto).

Se han consultado otros estudios en los que se define diferentes modos de tratar la evaluación de los OA, como es en Instrumento de Valoración de la Calidad de los OA de la UAA [7], la propuesta de un formato para evaluación de los OA [8], entre otros. Merlot es el único repositorio que realiza una evaluación de la calidad de los OA, donde almacena y muestra una lista con el ranking de los objetos evaluados [9].

De igual modo se continúa el desarrollo de los OA, donde las instituciones educacionales alrededor del mundo han enfocado sus investigaciones en iniciativas que aporten descripciones semánticas para la gestión de OA dentro de los repositorios. Por ejemplo el desarrollo de una arquitectura basada en ontologías para recuperar información relevante para los OA [10]; el proyecto elSEM (Sistemas de e-learning estandarizados basados en tecnología de Web Semántica) de la Universidad de Alcalá [11]; el proyecto Open Source Luisa [12], con el objetivo de enriquecer y hacer más flexibles los procesos de aprendizaje utilizando ontologías, entre otros.

Sin embargo hasta la actualidad no existe un consenso de los parámetros necesarios para la definición de un modelo de calidad estándar para evaluar los OA. Es decir cada organización define sus propios parámetros de calidad, donde se obvian otros que deberían tenerse en cuenta para alcanzar un nivel aceptable de calidad. El modo de evaluación que se utiliza actualmente en la práctica, es demasiado restrictivo, debido a que el tiempo es un obstáculo para la evaluación de los expertos en cada nuevo objeto de aprendizaje. Además se dificulta el proceso de actualización y mantenimiento de los repositorios de objetos de aprendizajes. Todo ello nos ha llevado a desarrollar la propuesta que se describe en la siguiente sección.

3. Propuesta de evaluación LOQEVAL

Despues de haberse analizado las deficiencias y dificultades encontradas en los estudios revisados y el uso de las ontologías en el desarrollo de los OA, en la sección 2, nos centramos en este trabajo a investigar y experimentar una nueva propuesta de evaluación de calidad de los OA haciendo uso de ontologías. Su objetivo principal es proporcionar nuevas herramientas a los evaluadores, que le faciliten evaluar la calidad de los OA de forma automática. Para ello se define un método de evaluación que se basa en 3 procesos desarrollados durante el ciclo de vida del OA. La descripción de cada uno de los procesos se expone a continuación:

1. Evaluación del OA durante su diseño.
   El primer proceso de evaluación se realiza durante el período de diseño del OA, a partir de un conjunto de parámetros predefinidos por el administrador del repositorio en el que se va a utilizar. Por ejemplo, el grado de confiabilidad de la fuente, el número de relaciones con otros OA, entre otros. A cada uno de los parámetros (f_j) utilizados se le asigna un peso (w_j) en función de la relevancia inicial que se les otorgue. De este proceso se obtiene un primer valor llamado calidad inicial (Q_i) del OA, que se describe como $Q_i = \sum_{j=1}^{n} (w_j \cdot f_j)$ para cada OA.
2. Evaluación del OA durante su vida útil.
   En una segundo proceso se realiza una evaluación del OA desde el punto de vista del usuario del OA, es decir, del uso del propio OA. Este proceso se desarrolla a partir de parámetros definidos previamente por el administrador del repositorio, por ejemplo, la reusabilidad del OA, donde se consideran el contexto en el que se encuentra el usuario y para el cual fue diseñado el OA, conjuntamente con los criterios positivos y negativos de los propios usuarios que utilizan los OAs en el repositorio. Estos parámetros se ajustan en función de las necesidades del repositorio. El resultado de este proceso se denominará calidad social (Q_s) de un OA.
3. Comparación de los resultados de ambos procesos y actualización de los parámetros del primer proceso.
   En esta último proceso, para cada OA se comparan los resultados del primer proceso de evaluación (Q_i) con los obtenidos en el segundo proceso de evaluacií (Q_s). El objetivo principal es minimizar la diferencia $|Q_i - Q_s|$, de forma que se consiga obtener un proceso de evaluación inicial que aproxime en gran medida la calidad del OA antes de ser utilizado. Este proceso inicial facilitará el uso de OAs recién introducidos en un repositorio de OAs. Conseguir que la diferencia entre ambos valores Q se reduzca implica la actualización tanto de los parámetros de evaluación como del peso asignado

a cada uno de ellos dentro del proceso inicial. Estos elementos serán actualizados de forma iterativa, cada vez que se utilicen los OAs, de modo que se obtendrá de forma gradual un proceso de evaluación más fiable ajustado al contexto del repositorio.

Para el desarrollo del proceso de evaluación con información se contará con información tanto del autor como del objeto de aprendizaje, conocimiento que será obtenido de dos ontologías, una ontología del modelo de autor y una del modelo de OA. La ontología del modelo del autor contendrá información relativa a las caracterìsticas que definen el perfil del autor, mientras que la ontología del modelo del OA refleja la estructura y características de los OAs.

A su vez es necesario disponer de una tercera ontología que modele los parámetros de calidad y las reglas de evaluación durante todos los procesos de la propuesta LOQEVAL.

4. Conclusiones

En el presente trabajo se define la propuesta LOQEVAL. A través de tres procesos diseñados, se trata de gestionar la evaluación de la calidad de los OAs, haciendo uso de ontologías. Se pretende obtener un modelo estándar de evaluación para OAs a partir del contexto en que se utilice. De este modo se contará con repositorios de objetos de aprendizajes con un nivel de calidad determinado.

Se plantea como trabajo futuro: i) diseñar e implementar computacionalmente la ontología de parámetros de evaluación; ii) implementar y validar los procesos de evaluación propuestos; y iii) definir un modelo estándar para la evaluación de los objetos de aprendizajes.

Referencias

1. Lopez, C.: Los Repositorios de Objetos de Aprendizaje como soporte a un entorno elearning. PhD thesis, Universidad de Salamanca (2005)
2. Moral, E.: Wikis, folksonom´ıas y webquests: trabajo colaborativo a través de objetos de aprendizaje. EDUTEC2006 (2006)
3. Nesbit, Belfer, O.L.: Learning Object Review Instrument (LORI). (2003)
4. Nesbit, Belfer, V.: A convergent participation model for evaluation of learning objects. Canadian Journal of Learning and Technology (2002)
5. Morales, Moreira, R.B.: Units of learning quality evaluation. SPDECE 2004, Design, Evaluation and Description of Reusable Learning Contents (2004)
6. Cabezuelo, A.S., Beardo, J.M.D.: Towards a model of quality for learning objects. In: ICALT '04: Proceedings of the IEEE International Conference on Advanced Learning Technologies, Washington, DC, USA, IEEE Computer Society (2004) 822–824
7. Velázquez, Muoz, A.: Aspectos de la calidad de objetos de aprendizaje en el metadato de lom. Virtual educa Brasil (2007)
8. Ruiz, Arteaga, R.: Evaluación de objetos de aprendizajes a través del aseguramiento de competencias educativas. Virtual educa Brasil (2007)
9. Merlot: Multimedia educational resource for learning and online teaching
10. Jovanovic, J., Knight, C., Gasevic, D., Richards, G.: Learning object context on the semantic web. In: ICALT '06: Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies, Washington, DC, USA, IEEE Computer Society (2006) 669–673
11. Carrion, J.S., Gordo, E.G., Sanchez-Alonso, S.: Semantic learning object repositories. International Journal of Continuing Engineering Education and Life Long Learning (2007)
12. Giorgini, P.: Leraning content management system using innovative semantic web services architecture (2008)

Semántica y sintáctica

Spanish-Basque SMT system: statistical translation into an agglutinative language

Díaz de Ilarraza, G. Labaka and K. Sarasola

Euskal Herriko Univertsitatea/Universidad del País Vasco

jipdisaa@ehu.es , jiblaing@ehu.es , jipsagak@ehu.es

In this paper we present the work done for adapting a baseline SMT system to carry out the translation into a morphologically-rich agglutinative language such as Basque. In translation from Spanish to Basque, some Spanish words, like prepositions or articles, correspond to Basque suffixes, and, in case of ellipsis, more than one of those suffix can be added to the same word. In this way, based on the Basque lemma *'etxe' /house/* we can generate *'etxeko' /of the house/, 'etxekoa' /the one of the house/, 'etxekoarengana' /towards the one of the house/* and so on.

Besides, Basque has another characteristic, it is a low-density language and there are less corpora available in comparison with other languages more widely used, like Spanish, English, or Chinese. For instance, the parallel corpus available for this work is 1M word for Spanish and 800K for Basque, smaller than the corpora usually used on public evaluation campaigns like NIST.

In order to deal with the problems presented above, we have split the Basque words on the lemma and some tags which represents the morphological information expressed on their inflection. In this way, we replaced the word *'etxekoa'* by *'etxe <IZE> <ARR> + <S> <GEN> + <S> <ABS>' /house<noun><common> +<sing><genitive> +<sing><absolutive>/*. Dividing the Basque word in this way, we expect to reduce the sparseness produced by the agglutinative being of Basque and the small amount of training data.

Working at the morpheme level, the output of our SMT system is a sequence of morphemes. So to produce the final Basque text, we need to recover the words from this sequence of morphemes, so the output of the SMT system is post-processed to produce the final Basque translation. In order to incorporate a word level language model, we use a n-best list which is reranked according to this language model.

We have try different segmentation levels, from the fine-grained segmentation reported by the analyzer to the most coarse-grained where all suffixes of one lemma are packed in a unique token. Those differences in segmentation have a significant impact on translation quality. Best results (obtined by the coarse-grained segmentation) significally outperforms word-level translation on all automatic metrics used (BLEU, NIST, WER and PER).

Estructura argumental nominal

Aina Peris Morant

CLiC-UB (Universitat de Barcelona)

aina.peris@ub.edu

El objetivo de mi trabajo de investigación consiste en el estudio lingüístico de la estructura interna de los sintagmas nominales (SNs, en adelante) que tienen como núcleo nombres con capacidad argumental. De este estudio obtendremos una clasificación nominal de este tipo de nombres en función de sus rasgos morfosintácticos y semánticos que se representará en el lexicón nominal AnCora-Nom. Este lexicón será la base de la anotación de estos SNs en el corpus AnCora.

Como punto de partida nos hemos centrado en los sustantivos deverbales ya que se considera que heredan la estructura argumental de los verbos. Para este tipo de nombres se ha realizado una propuesta inicial de entrada léxica donde se especifican los siguientes atributos: lema, sentido nominal (synset de WordNet), tipo de nombre (deverbal, deadjetival o relacional), subtipo denotativo (eventivo, resultativo o no marcado, para los nombres deverbales), categoría de palabra de la que derivan (verbos, por ahora), el verbo en cuestión, clase semántica de dicho verbo, complementos del nombre y sus respectivas funciones sintácticas con el argumento y papel temático correspondiente, el tipo de determinante y la pluralidad.

En un primer estudio se ha analizado el subtipo denotativo del sustantivo deverbal ya que es un punto controvertido en la bibliografía. Los autores reconocen la diferencia entre sustantivos que denotan un evento (1a) y un resultado (1b).

(1a) Lo que condujo a **su combinación para formar el complejo n-molecular dador aceptor**.

(1b) **De dicha combinación** nace una criatura con características propias.

En la siguiente tabla se muestran los distintos criterios lingüísticos usados por diferentes autores para diferenciar entre ambos tipos denotativos:

Criterios Lingüísticos	Grimshaw (1990)	Alexiadou (2001)	Picallo (1999)	Alonso (2004)	Badia (2002)
Clase Verbal	-	+	+	-	+
Pluralización	+	-	-	+	-
Tipo de Determinante	+	-	+	+	-
Obligatoriedad del argumento interno	+	-	+	-	-
Modificadores del agente	+	-		-	-
Modificadores AspectualesTemporales	+	+	+	-	-
Poseedores vs. Argumentos	+	+	-	-	-
Estructuras de Control	+	-	+	-	-
Predicado Verbal + SN	+	-	+	-	+
Preposición + Agente	-	-	+	-	+

Hasta el momento se han analizado unos 400 sustantivos deverbales de un subconjunto de 100.000 palabras de AnCora-Es. Este análisis nos ha permitido concluir que:

1. Hay seis criterios lingüísticos que parecen más discriminadores: la clase verbal de la que deriva el sustantivo, la obligatoriedad del argumento interno, la capacidad de pluralización, el tipo de determinante, la interpretación argumental de los adjetivos temáticos y la preposición que introduce el agente.
2. No siempre es posible distinguir entre sustantivos eventivos y resultativos. Si el contexto está infraespecificado, ambas lecturas parecen ser recuperables. Así pues, la especificación del contexto un factor a tener en cuenta para distinguir ambas denotaciones.

Nuestro objetivo es seguir analizando los diferentes aspectos de la estructura léxica nominal.

Resolución de expresiones anafóricas en textos biomédicos.

Sergio Aparicio Escribano, Isabel Segura Bedmar

Universidad Carlos III de Madrid

.

La resolución de la anáfora es una tarea fundamental para comprender el mensaje de un texto. En la literatura biomédica puede contribuir a mejorar los sistemas de extracción de información, en particular, la extracción de interacciones entre fármacos. En este documento se presenta una aproximación más a la resolución de anáforas de tipo pronominal y de expresiones definidas en textos biomédicos en inglés.

La anáfora es un tipo de deixis que desempeñan ciertas expresiones para recoger el significado de una parte del discurso ya emitida. La parte anterior del discurso se denomina antecedente de la expresión anafórica. Es necesario identificar el antecedente para poder interpretar el mensaje que aporta la expresión anafórica. En el siguiente ejemplo la primera expresión entre corchetes es el antecedente de la expresión anafórica (segunda expresión entre corchetes):

> ***[Levofloxacin]***$_i$*, a fluoroquinolone, is one of the most commonly prescribed antibiotics in clinical practice. Several case reports have indicated that* ***[this drug]***$_i$ *may significantly potentate the anticoagulation effect of warfarin.*

La resolución de la anáfora consiste en detectar expresiones anafóricas y posteriormente identificar su o sus antecedentes. En la literatura biomédica, los tipos más comunes de anáfora son principalmente dos: las anáforas pronominales y las anáforas en expresiones definidas.

Las anáforas pronominales son aquellas en las que la expresión anafórica es un pronombre personal o un posesivo. En la literatura biomédica no se consideran expresiones anafóricas pronombres personales como *I*, *you*, *he*, *she* o *we*:

> ***[Aspirin]***$_i$ *is an useful salicylate. Several case reports have indicated that* ***[it]***$_i$ *increases the effect and toxicity of methotrexate.*

En la búsqueda de expresiones anafóricas de tipo pronominal, es muy importante reconocer los pronombres pleonásticos y excluirlos de la selección. Un pronombre pleonástico es un pronombre no anafórico y por lo tanto carece de antecedente. El siguiente ejemplo de pronombre pleonástico es común en la literatura biomédica:

> ***It*** *is not recommended that [..]*

El otro tipo de anáfora más común en la literatura biomédica es la anáfora en expresiones definidas. Este tipo de anáfora es aquella que se aparece en estructuras que

comienzan con el artículo determinado *the*. También se consideran de éste tipo expresiones con las palabras *each, both, that, those, this,* y *these*.

- ***[Aspirin]$_i$*** *is a salicylate drug.* ***[The medicine]$_i$*** *is used as an analgesic.*
- *The possibility that* ***[carbamazepine]$_i$*** *might increase the clearance of* ***[escitalopram]$_i$*** *should be considered if* ***[the two drugs]$_i$*** *are coadministered.*

Cada expresión anafórica puede tener uno o varios antecedentes que es necesario identificar. Para llevar a cabo esta identificación se asignan pesos a cada uno de los candidatos a antecedente para valorarlos. Esta valoración se lleva a cabo mediante varios factores:

- Distancia: descarta candidatos demasiado lejanos a la expresión anafórica y asigna mayor puntuación a los candidatos más cercanos.
- Morfología: hace referencia a la estructura interna de las palabras. Cuanto mayor parecido exista entre el núcleo de la expresión anafórica y el núcleo del candidato a antecedente, mayor prioridad se le dará a dicho antecedente. Esta similitud puede llevarse a cabo mediante la función LCS (*Longest Common Subsequence*). Este factor solo se emplea para las anáforas en estructuras definidas.
- Semántica: se valoran más a aquellos candidatos que compartan la clase semántica con la expresión anafórica. Este factor solo se lleva a cabo en el caso de las anáforas en estructuras definidas.
- Concordancia numérica: elimina aquellos candidatos que no concuerdan en número con la anáfora.

Finalmente, tras valorar los candidatos a antecedente, es necesario seleccionar los más adecuados:

- En el caso en el que una expresión anafórica indique el número concreto de antecedentes que requiere (*the two drugs*) se seleccionan aquellos n antecedentes con mayor valoración.
- En el caso de que el número de antecedentes no sea concreto (*those drugs*) se seleccionan todos aquellos antecedentes que estén valorados por encima de un peso mínimo denominado umbral mínimo.

Este algoritmo ha sido evaluado con 20 ficheros del corpus Drugner (un corpus de interacciones farmacológicas procesado con MetaMap y etiquetado con clases semánticas de UMLS). El número total de anáforas del corpus de evaluación es de 120. Los ficheros del corpus no estaban perfectamente etiquetados por lo que los resultados de la evaluación no son demasiado precisos.

Precision	**Recall**	**F-Measure**
63.50%	72.50%	67.70%

Análisis sintáctico profundo en FreeLing

Irene Castellón, Nevena Tinkova
Universitat de Barcelona
{icastellon, nevenatinkova}@ub.edu

Jordi Carrera, Marina Lloberes, Lluís Padró
Universitat Politècnica de Catalunya
{jcarrera, mlloberes, padro}@lsi.upc.edu

Resumen: La versión del analizador TXALA de FreeLing y las gramáticas de dependencias que presentamos suponen un incremento del potencial del analizador y de la expresividad de las gramáticas con el objetivo de abarcar un mayor número de fenómenos lingüísticos del castellano, el catalán y el inglés.

Palabras clave: Análisis sintáctico automático, análisis sintáctico profundo, gramática de dependencias.

Abstract: FreeLing deep parser, TXALA, and dependency supported grammars are an extension of a previous version. These FreeLing tools that we present include a new parser and grammar options to solve more linguistic phenomena in Spanish, Catalan and English grammars.

Kewords: Automatic parsing, deep parsing, dependency grammar.

1.Introducción

El analizador TXALA y las gramáticas de dependencias que presentamos suponen una ampliación de la primera versión (Atserias, J. et al., 2005) en cuanto a la expresividad y el número de fenómenos que tratan. Se trata de un módulo de FreeLing (Atserias, J. et al, 2006) y se incluyen como una de las tareas del proyecto KNOW (TIN200615049C03, Ministerio de Industria). FreeLing se presenta como una librería de herramientas de PLN para las lenguas del Estado Español (castellano, catalán, vasco y, además, inglés e italiano) desarrollada bajo la licencia GNU Lesser General Public License (LGPL).

Si se observa la cobertura de los analizadores sintácticos automáticos en función de las lenguas que tratan, son escasos los que emplean como modelo el castellano (Bick, E., 2006; Ferrández, A. et al., 2006; Marimon, M. et al, 2007; Tapanainen, P. y Järvinen, T., 1997), el catalán (Alsina,À. et al., 2002; Atserias, J., 2005) y el vasco (Bengoetxea, K. Y Gojenola, K., 2007). Al mismo tiempo, aunque el formalismo de dependencias ha sido implementado en PLN (By, T., 2004), existen pocos analizadores de dependencias para estas lenguas, DILUCT (Gelbukh, A. et al., 2005) y Connexor(http://www.connexor.com).

2. El analizador de dependencias TXALA

El objetivo principal del analizador es proporcionar análisis sintácticos profundos y robustos, es decir, resolver las agrupaciones de los nodos de la estructura sintáctica y ofrecer siempre un análisis para toda estructura. Para ello, TXALA emplea tres operaciones.

Puesto que la información lingüística de entrada son los árboles sintácticos generados por el analizador superficial de FreeLing (Atserias, J. et al., 1998), TACAT, en primer lugar, se completan los análisis parciales mediante reglas definidas manualmente. En ellas, se define la prioridad en la cual TXALA agrupará cada par de árboles, y unas restricciones que se refieren al contexto, al control del número de aplicación de reglas y a la información morfológica (PoS), léxica (lema y forma) y sobre clases de palabras preestablecidas.

Simultáneamente, se transforma el árbol sintáctico en un árbol de dependencias a través de la determinación en uno de los campos de las reglas del núcleo del árbol que actuará como el nodo de agrupación del resto de nodos inferiores del árbol.

Una vez llevada a cabo la transformación del árbol, se asigna cada dependencia con la función sintáctica que le corresponda. Esta tarea la realizan un conjunto de reglas escritas también manualmente donde se expresa la etiqueta de la relación de dependencia, la prioridad y un conjunto de condiciones que atañen a la posición en el árbol y a la información léxica (lema) y semántica (clases preestablecidas, *synsets* de WordNet -sinónimos e hiperónimos- y atributos de Top Ontology) de los nodos que constituyen la dependencia.

3. Las gramáticas de dependencias de TXALA

TXALA incluye las gramáticas de dependencias para el castellano (4349 reglas), el catalán (2900 reglas) y el inglés (1857 reglas). Gran parte de las reglas se concentran en la tarea de compleción y transformación del árbol (3777 para el castellano, 2362 para el catalán y 1606 para el inglés) por el hecho que son muy diversos y complejos los fenómenos sintácticos en este nivel de análisis.

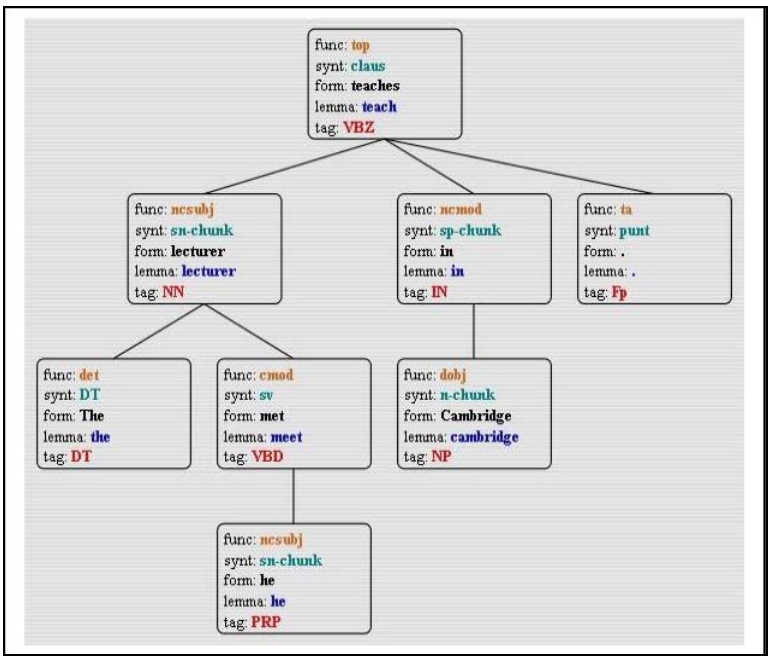

Fig. 1. Análisis de la frase
The lecturer he met teaches in Cambridge.

La principal limitación con que se encuentran los analizadores sintácticos automáticos es la resolución de las ambigüedades propias del lenguaje natural. Es bien conocida la dificultad del tratamiento del sintagma preposicional ya que puede ser anidado en diferentes posiciones del árbol (argumento verbal, modificador nominal o adjetival, etc.). Como consecuencia, en las gramáticas de TXALA se han llevado a cabo una serie de estrategias que varían en función de la lengua. Concretamente, en castellano y en catalán, se ha acotado la agrupación del sintagma preposicional mediante, la prioridad, la restricción del contexto, la definición de clases de verbos y de sustantivos que exigen preposición y, en algunos casos, mediante la especificación de información léxica.

Por otro lado, aunque no menos relevante, las subordinadas introducidas por conjunción con frecuencia son problemáticas puesto que un analizador sintáctico no las puede distinguir de algunas oraciones interrogativas si no se facilita información adicional. Pero, además, el inglés presenta, en este campo, una complicación añadida debido a la posibilidad de omitir la partícula de la subordinada. La gramática de dependencias trata este fenómeno con bastante eficacia gracias a la prioridad y al contexto (Fig. 1).

A pesar de que los mayores esfuerzos se empleen en las dos primeras operaciones de TXALA, la tarea de etiquetar las dependencias conlleva en mayor parte el hecho de tratar el reconocimiento entre argumentos y adjuntos, por un lado, y la determinación del tipo de argumento, por el otro lado. En este último caso, las gramáticas del castellano y del catalán recurren a clases de verbos enlazados a las reglas como módulos externos (v. Fig 2).

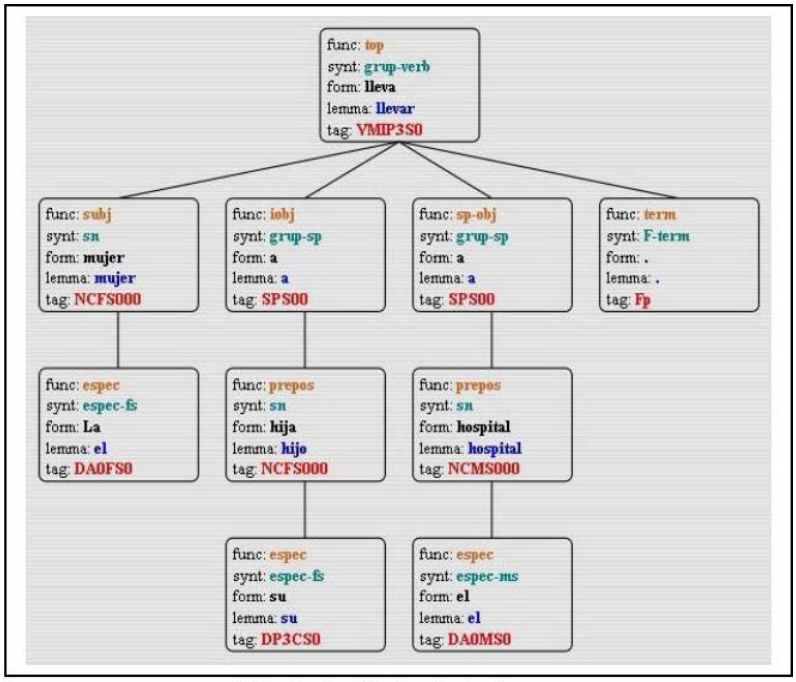

Fig 2. Análisis de la frase

La mujer lleva a sufija al hospital.

4. Trabajo futuro

No obstante las gramáticas de dependencias de FreeLing tratan un amplio repertorio de fenómenos, es necesario plasmar en corpus lingüísticos la cobertura y robustez de ellas. Por este motivo, estamos iniciando un proceso de evaluación de las tres gramáticas y de los recursos que utilizan que tiene como objetivo abarcar tanto la vertiente cuantitativa como la cualitativa. De modo que los resultados que proveerá la evaluación indicarán las directrices para desarrollar la futura versión de TXALA y de las gramáticas.

Referencias

Alsina, À., T. Badia, G. Boleda, S. Bott, Á. Gil, M. Quixal, y O. Valentín. 2002. CATCG: Un sistema de análisis morfosintáctico para el catalán. En Procesamiento del Lenguage Natural, n. 29, pp. 309310.

Atserias, J., I. Castellón y M. Civit. 1998. Syntactic Parsing of Unrestricted Spanish Text. First International Conference on Language Resources and Evaluation (LREC'98).

Atserias, J., E. Comelles y A. Mayor. 2005. TXALA un analizador libre de dependencias para el castellano. En Procesamiento del Lenguaje Natural, n. 35, pp. 455-456.

Atserias, J., B. Casas, E. Comelles, M. González, L. Padróy M. Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. En Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06).

Bengoetxea, K. y K. Gojenola. 2007. Desarrollo de un analizador sintáctico estadístico basado en dependencias para el euskera. En Procesamiento del Lenguaje Natural, n. 39, pp. 5-12.

Bick, E. 2006. A Constraint GrammarBased Parser for Spanish. En Proceedings of TIL 2006. 4th Workshop on Information and Human Language Technology.

By, T. 2004. English dependency grammar. En RADG 2004.

Ferrández, A., M. Palomar y L. Moreno. 2000. “Slot Unification Grammar and anaphora resolution”. En Recent Advances in Natural Language Processing. Nicolas Nicolov y Ruslan Mitkov (eds). John Benjamins: Amsterdam & Philadelphia, pp. 155-166.

Gelbukh, A., H. Calvo y S. Torres. 2005. Transforming a constituency Treebank into a dependency treebank. En Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN’2005).

Marimon, M., N. Bel y N. Seghezzi. 2007. Test Suite Construction for a Spanish Grammar. En T. Holloway King y E.M. Bender (eds.). Proceedings of the Grammar Engineering Across Frameworks (GEAF2007) Workshop"CSLI Studies in Computational LinguisticsONLINE”, pp. 250264.

Tapanainen, P y T. Järvinen. 1997. A nonprojective dependency parser. En Proceedings of the 9th International Joint Conference.

Domain Adaptation for Supervised Word Sense Disambiguation

Oier Lopez de Lacalle and Eneko Agirre

IXA Taldea (University of the Basque Country)

In this paper we explore robustness and domain adaptation issues for Word Sense Disambiguation (WSD) using Singular Value Decomposition (SVD) and unlabeled data in order to extract more reliable features. We focus on the semi-supervised and supervised domain adaptation scenarios. In the first scenario we train on the source corpus (out of domain data set) and test on the target corpus, and try to improve results using unlabeled data. Our method yields up to 16.3% error reduction compared to state-of-the-art systems, being the first to report successful semi-supervised domain adaptation. Surprisingly the improvement comes from the use of unlabeled data from the source corpus, and not from the target corpora, meaning that we get robustness rather than domain adaptation.

In supervised domain adaptation scenario the WSD systems are trained on source and target domain data. We show that using source examples from the British National Corpora (BNC), we are able to adapt the WSD system with examples from two target domains (Sports and Finances), obtaining up to 22% error reduction when compared with the performance on the target domain alone, and improving over a state-of-the-art domain adaptation algorithm. We also show that as little as 40% of the target data is sufficient to adapt the WSD system trained on BNC. These results are remarkable given the scarce positive results on supervised domain adaptation for WSD. The key for success is the use of unlabeled data with SVD, and the combination of kernels using SVM. We also will provide further analisys of the results.

Using Word Sense Disambiguation for (Cross Lingual) Information Retrieval

Arantxa Otegi, Eneko Agirre, German Rigau

IXA NLP Group - University of the Basque Country
Donostia, Basque Country.

arantza.otegi@ehu.es

This contribution describes the participation of the IXA NLP group at the CLEF 2008 Robust-WSD Task. This is our first time at CLEF, and we participated at both the monolingual (English) and the bilingual (Spanish to English) subtasks. We tried several query and document expansion and translation strategies, with and without the use of the word sense disambiguation results provided by the organizers. All expansions and translations were done using the English and Spanish wordnets as provided by the organizers and no other resource was used. We used Indri as the search engine, which we tuned in the training part. Our main goal was to improve (Cross Lingual) Information Retrieval results using WSD information, and we attained improvements in both mono and bilingual subtasks, although the improvement was only significant for the bilingual subtask. As a secondary goal, our best systems ranked 4th overall and 3rd overall in the monolingual and bilingual subtasks, respectively.

Linking WordNet to FrameNet by using a knowledge-base Word Sense Disambiguation algorithm

Egoitz Laparra and German Rigau

IXA NLP Group, UPV/EHU Donostia, Basque Country

1 Introduction

Models of lexical semantics are core paradigms in most NLP applications, such as dialogue, information extraction and document understanding. Unfortunately, the coverage of currently available resources is still unsatisfactory. Following the line of previous works [1], [2], [3], we present a new approach for extending the FrameNet coverage using a knowledge-based Word Sense Disambiguation algorithm for linking each lexical-unit from FrameNet[4] to a synset from WordNet [5].

1.1WordNet

WordNet[1] is a large dictionary whose basic unit is the synset, i.e. an equivalence class of word senses under the synonymy relation representing a concept. Synsets are organized hierarchically using the is-a relation. WordNet covers the vast majority of nouns, verbs, adjectives and adverbs from the English language. WordNet also includes an impressive number of semantic relations defined across concepts, including hyperonymy/hyponymy /IS_A), meronymy/holonomy (HAS_A), antonomy, entailment, etc.

1.2FrameNet

FrameNet[2] is a medium-sized lexical database that lists descriptions of English words in Fillmore paradigm of Frame Semantics [6]. In this framework, the relations between predicates, or in FrameNet terminology, target words, and their arguments are described by means of semantic frames. A frame can intuitively be thought of as a template that defines a set of slots, frame elements, that represent parts of the conceptual structure of a predicate and correspond to prototypical participants or properties.

[1] http://wordnet.princeton.edu/

[2] http://framenet.icsi.berkeley.edu/

The initial versions of FrameNet focused on describing situations and events, i.e. typically verbs and their nominalizations.

Currently, however, FrameNet defines frames for a wider range of semantic relations, such as between nouns and their modifiers. FrameNet frames typically describe events, states, properties, or objects. The different senses of a word in FrameNet are represented in different frames.

2 FrameNet lexical-unit disambiguation

2.1 1SSI-Dijkstra

We used a wide-coverage knowledge-based Word Sense Disambiguation (WSD) algorithm for linkingeach lexical-unit from FrameNet to a synset from WordNet. Since the lexical-units of a frame are closely semantically related (for instance, among others therapeutic.a, therapist.n, therapy.n, treat.v, treatment.n are assigned to the frame Cure), we processed each frame by applying the WSD algorithm to its associated lexical-units. We have used the SSI-Dijkstra algorithm [7]. This is a version of the Structural Semantic Interconnections algorithm (SSI), a knowledge-based iterative approach to Word Sense Disambiguation [8]. The Dijkstra algorithm is a greedy algorithm for computing the shortest path distance between one node an the rest of nodes of a graph. In that way, the SSI-Dijkstra algorithm can compute very efficiently the shortest distance between any two given nodes of a graph that contains the relations between synsets from Word-Net. The main drawback of this algorithm is that itneeds at least a monosemous word in the set of words to be disambiguated. For overcome this problem, we have implemented four different versions of the SSI-Dijstra algorithm that can work even if there are only polysemous words in the set of words to be disambiguated.

2.2Evaluation

We have evaluated the outputs of our different versions of the algorithm using a gold-standard provided by Sara Tonelli from the Fondazione Bruno Kessler at Trento. This gold-standard consists of a set of 123 lexical-units manually assigned to a synset from WordNet 1.6. Table 1 shows the accuracy figures of the best system.

Accuracy	**(%)**
Nouns	75.6
Verbs	54.4
Adjectives	85.7
Total	**66.1**

Table1.

3 Conclusions

In this work, we have presented an approach for mapping Framenet to WordNet using a general purpose WSD algorithm. Currently, we have developed four different versions that improve the initial SSI-Dijktra algorithm. These versions obtained encouraging results when frames are composed mainly by nouns or adjetives, however, relatively poor results are obtained when the frame is mainly composed by verbs.

References

1. Burchardt,A., Erk,K., Frank,A. : A WordNet Detour to FrameNet. In: Proceedings of the GLDV2005 GermaNet II Workshop, Bonn, Germany (2005) 408–421
2. Shi,L., Mihalcea,R.: Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. Computational Linguistics and Intelligent Text Processing(2005) 100–111
3. Johansson,R., Nugues,P.: Using WordNet to extend FrameNet coverage. In: Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA, Tartu, Estonia (May242007)
4. Baker,C., Fillmore,C., Lowe,J.: The Berkeley framenet project. In: COLING/ACL'98, Montreal, Canada(1997)
5. Fellbaum,C., ed.: WordNet. An Electronic Lexical Database. The MIT Press (1998)
6. Fillmore,C.: Frame semantics and the nature of language. In Annals of the New York Academy ofSciences: Conference on the Origin and Development of Language and Speech 28 (1976)20–32
7. Cuadros,M., Rigau,G.: KnowNet: Building a Large Net of Knowledge from the Web. In: Proceedings of COLING.(2008)
8. Navigli,R.,Velardi,P.: Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27 (7) (2005) 1063–1074

Aplicación de los Roles Semánticos a la identificación de Expresiones Temporales *

H. Llorens, E. Saquete, B. Navarro

Grupo de Procesamiento del Lenguaje Natural y Sistemas Informáticos
Universidad de Alicante, España

1. Introducción

En los últimos años, la identificación automática de expresiones temporales (ETs), eventos y sus relaciones sobre texto en lenguaje natural está teniendo una gran relevancia en el área del procesamiento del lenguaje natural (PLN). Prueba de ello es el número de conferencias que ha generado sobre su aplicación a la extracción de información (EI) y a la búsqueda de respuestas (BR) (TERQAS 2002, TANGO 2003, Dagstuhl 2005), así como el foro de evaluación TempEval2007 (SemEval2007).

La necesidad de anotar ETs, eventos y sus relaciones desembocó en la creación de un lenguaje de marcado XML, el lenguaje TimeML [4] que es considerado un estándar de-facto para esta tarea. Junto a la primera versión de TimeML se desarrolló el corpus TimeBank [5]. Se trata de un corpus en inglés anotado en TimeML, que a día de hoy, y tras varias revisiones, es considerado como un corpus de referencia del lenguaje.

Por otro lado, los roles semánticos (RS) han despertando gran interés en su aplicación a distintas áreas del PLN [1,2]. Para cada predicado de una frase, los RS identifican todos los constituyentes, determinando sus roles (agente, paciente,etc.) y sus adjuntos (locativo, temporal, etc.). De esta manera, el rol temporal representa *cuándo* tuvo lugar un evento representado por el verbo de una frase, lo que indica que podrían ser útiles en la identificación de ETs en los textos.

Tomando estas premisas como punto de partida, esta investigación pretende estudiar en qué medida es efectiva la aplicación de los RS para la identificación de ETs (TIMEX3) siguiendo el estándar TimeML. Para ello se presenta y evalúa un sistema capaz de identificar ETs sobre textos planos haciendo uso de los RS.

2. Implementación

Con el objetivo de estudiar la aplicación de los RS a la identificación de ETs siguiendo las indicaciones del TimeML se ha desarrollado un sistema capaz de tomar una entrada en texto plano y, usando la información de los RS, convertirla

* Este artículo ha sido financiado por el MICINN, proyecto TEXT-MESS TIN-2006-15265-C06-01 donde Héctor Llorens dispone de una beca FPI (BES-2007-16256)

en un texto XML etiquetado según las especificaciones del TimeML. Para el etiquetado del texto con RS se ha utilizado la herramienta desarrollada por el grupo CCG de la Universidad de Illinois[3]. Han sido implementadas dos versiones:

2.1. Baseline

En primer lugar se desarrolló un sistema básico (Baseline) que etiqueta todos los roles temporales como ETs. Puede darse el caso de que en una frase haya más de un verbo y que cada uno tenga un rol temporal distinto. En caso de que dos roles temporales se solapen nuestro sistema elegirá el que tenga menos palabras. Ésto es así puesto que en la mayoría de los casos se trata de una subordinación, como se muestra a continuación.

```
RS para estar:
[Estaba V] [en la terraza AM-LOC] [cuando me llamó ayer AM-TMP].
RS para llamar:
[Estaba en la terraza *] [cuando R-AM-TMP] [me A2] [llamó V] [ayer AM-TMP].
Resolución: Estaba en la terraza cuando me llamó <TIMEX3>ayer</TIMEX3>.
```

2.2. Adapted

Tras la prueba del sistema Baseline se advirtió que los roles temporales no se adaptaban exactamente a las especificaciones del TimeML. El problema más importante era que los elementos de la frase marcados como roles temporales incluían las señales temporales. Según el estándar TimeML [4], las señales temporales deben ser etiquetadas como SIGNAL y son elementos textuales que hacen explicita la relación entre dos entidades temporales, dos eventos o una entidad temporal y un evento.

```
RS: She was born [in 1981 AM-TMP].
Baseline: She was born <TIMEX3>in 1981</TIMEX3>.
TimeML: She was born <SIGNAL>in</SIGNAL> <TIMEX3>1981</TIMEX3>.
```

Otro problema es que la subordinación temporal se marca como AM-TMP. En la mayoría de los casos se puede detectar porque abarca un mayor número de palabras que las ETs y porque suele contener un verbo mientras que las ETs no (ej.: before any attempts at seizure are made).

Finalmente, un problema intrínseco de los RS es que no pueden anotarse en frases que no tengan verbo, dado que el verbo es el elemento central de dicha anotación (ej.: títulos, paréntesis).

Hechas estas observaciones se implementó una versión del sistema adaptada (Adapted) a las especificaciones del TimeML de la siguiente manera:

- **Omisión de la preposición inicial**: Se omiten todas las preposiciones iniciales de los roles temporales al etiquetarlos como ETs utilizando una lista de preposiciones temporales (SIGNAL) y la categoría léxica de las palabras.

- **Omisión de subordinación temporal**: Se omiten todas las ETs contenidas en una subordinación temporal que tengan verbo o una extensión superior a 6 palabras.
- **Etiquetado básico de ETs para frases sin verbo**: Utilizando un etiquetador simple de ETs basado en reglas el sistema es capaz de anotar ETs básicas como fechas, días de la semana, etc.

3. Evaluación

Los sistemas desarrollados, Baseline y Adapted, han sido evaluados y comparados con los sistemas GUTime1[1] y TARSQUI Tool Kit [6].

3.1. Benchmark

Como corpus se ha utilizando la versión 1.2 del TimeBank[2] formada por 183 artículos periodísticos anotados según la especificación del TimeML 1.2.1[3]. Partiendo de los textos del corpus sin etiquetar se ha comparado el etiquetado de ETs (TIMEX3) de cada uno de los sistemas con el etiquetado original del corpus.

3.2. Resultados

En la tabla 1 se presentan los resultados obtenidos. Para cada sistema se indica los resultados Inexactos *I* y Exactos *E*. En los resultados *I* se consideran correctos aquellos casos dónde se ha identificado la ET pero no se ha delimitado de forma exacta. Las medidas utilizadas en el orden en que aparecen son: *pos* número total de ETs, *act* número identificado de Ets, *corr* correctas, *inco* incorrectas, *miss* no detectadas y *spur* falsos positivos. También se ha calculado la precisión *prec*, la cobertura *rec* y la medida *F1*.

Nuestro sistema Baseline obtiene un 63.7 % de F1 en la identificación inexacta, pero en la exacta cae hasta un 37.1 %. En cambio la versión adaptada alcanza un 81.5 % de F1 en la identificación inexacta y un 71.4 % en la exacta. Los sistemas del estado de la cuestión TTK y GUTime1 obtienen resultados alrededor de el 60 % de F1 en identificación inexacta y alrededor del 48 % en identificación exacta.

Comparando nuestros sistemas Baseline y Adaptado vemos que el Adaptado prácticamente dobla los resultados del Baseline lo cual indica que las simples mejoras introducidas afectan muy favorablemente a su eficacia. Por otro lado, comparando los resultados con los sistemas TTK y GUTime1 observamos que la versión Baseline obtiene unos resultados similares en la identificación inexacta pero más bajos en la exacta, sin embargo, la versión Adapted mejora sus resultados considerablemente en identificación exacta (+47.52 % en F1).

[1] http://complingone.georgetown.edu/~ linguist/
[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T08
[3] http://www.timeml.org/site/publications/specs.html

Test		Resultados								
Nombre	**T**	**pos**	**act**	**corr**	**inco**	**miss**	**spur**	**prec**	**rec**	**F**
Baseline	I	1441	1390	892	0	519	498	0.642	0.632	0.637
	E	1441	1390	519	373	519	498	0.373	0.368	**0.371**
Adapted	I	1441	1565	1212	0	199	353	0.774	0.859	0.815
	E	1441	1565	1062	150	199	353	0.679	0.753	**0.714**
TTK	I	1441	1244	836	0	575	408	0.672	0.592	0.630
	E	1441	1244	628	208	575	408	0.505	0.445	**0.473**
GUTime1	I	1441	1472	862	0	549	610	0.585	0.610	0.597
	E	1441	1472	699	163	549	610	0.474	0.495	**0.484**

Cuadro 1. Resultados de la identificación de ETs (TIMEX3)

4. Conclusiones y trabajo futuro

Este trabajo presenta un sistema basado en RS para la detección de ETs según las indicaciones del TimeML. De este sistema se han desarrollado dos versiones: (1) Baseline y (2) Adapted. Siendo Baseline una aplicación directa del rol temporal al etiquetado de ETs y la Adapted una adaptación de este etiquetado a las indicaciones del TimeML. Estos sistemas junto con dos sistemas del estado de la cuestión han sido evaluados utilizando el TimeBank como corpus. Los resultados obtenidos en el análisis confirman que los RS pueden ser de gran ayuda para la identificación de ETs, obteniendo, Adapted, mejores resultados que los sistemas del estado de la cuestión evaluados (+47.52 % mejora F1).

Como trabajo futuro se plantea extender el uso de los RS a la detección de señales, eventos y relaciones temporales, así como el estudio de su aplicación al castellano.

Referencias

1. D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.
2. P. Moreda, H. Llorens, E. Saquete, and M. Palomar. Automatic generalization of a qa answer extraction module based on semantic roles. In *AAI - IBERAMIA*, volume 5290 of *LNAI, LNCS*, pages 233–242. Springer, 2008.
3. V. Punyakanok, D. Roth, W.-t. Yih, D. Zimak, and Y. Tu. Semantic role labeling via generalized inference over classifiers. In *HLT-NAACL (CoNLL-2004)*, pages 130–133, MA, USA, 2004. ACL.
4. J. Pustejovsky, J. M. Castaño, R. Ingria, R. Saurí, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev. Timeml: Robust specification of event and temporal expressions in text. In M. T. Maybury, editor, *New Directions in Question Answering*, pages 28–34. AAAI Press, 2003.
5. J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. J. Gaizauskas, A. Setzer, D. R. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. The TIMEBANK corpus. In *Corpus Linguistics*, pages 647–656, 2003.
6. M. Verhagen, I. Mani, R. Saurí, R. Knippen, S. B. Jang, J. Littman, A. Rumshisky, J. Phillips, and J. Pustejovsky. Automating temporal annotation with tarsqi. In *ACL*, pages 81–84, NJ, USA, 2005. ACL.

Extraction of Temporal Semantics for Improving Web Search

María Teresa Vicente-Díez

Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganés, Madrid, Spain
tvicente@inf.uc3m.es

Abstract. Current Web-search engines can be improved through techniques that consider a temporal dimension both in query formulation and document extraction processes. An accurate recognition of temporal expressions in data sources must firstly be done. Such temporal information must be dealt with in an appropriated standard format that allows reasoning without ambiguity. We propose a temporal expressions recognition and normalization (TERN) system for contents in Spanish, which has been integrated into a Web-search engine prototype. The contribution of this system lies in its capability of taking into account the different ways of expressing time in the user queries as well as in the documents, independently of the format. This approach shows that the inclusion of temporal information management aptitudes to Web searching means considerable improvements.

Keywords: Temporal Web-search, Time Semantics Extraction.

1 Introduction

The amount of accessible information on the Web is daily increased. Insofar as this happens search engines become an essential tool for thousands of users. Therefore it is suitable that engines are capable of providing relevant information accurately to the search criteria. However, up-to-date Web search engines do not consider all semantic information from the contents that they handle when elaborating their results. A concrete case of semantic loss happens when embedded temporal information is not taken into account both in documents and in search criteria [1]. In fact, if a user introduces a question whose search terms include some temporal expression these are managed as independent terms, without taking advantage of subjacent semantic. However, when a system deals with temporal semantics, it must understand the nature of time and being capable to accept and reason with time-related facts [2]. Because of this lack, it is frequent to find that searching information related to a temporal expression, retrieved results only points to documents in which the same terms of the query are literally represented, discarding others that even mentioning relevant information, lacks the same representation. Moreover, current approaches of matching strings are not sufficient as relative times change based on the time of search.

The extraction and tagging of time expressions or *timexes* must be done both in the search terms and source collections. On the other hand, a mechanism for

normalization of extracted timexes is required in search engines, so their representation is standardized and their management is uniform [3].

2 System Architecture

This approach has been developed in two steps: in the first one a TERN system for Spanish was implemented [4] in order to detect and extract temporal information and semantics. The second one consists on the creation of a search engine integrating the previous system. It allows a proper semantic interpretation when formulating queries demanding a temporal information correspondence.

2.1 TERN System

The TERN system processes input data from documents identifying temporal expressions (recognition stage) that can be classified into several types, according to their way of definition and resolution [5]. Next, recognized expressions are managed and returned in a standard format that avoids semantic ambiguities in their retrieval (normalization stage). The reference date for relative temporal expressions resolution is taken from the documents (depending of the case, the creation date of the document or a date extracted from the context is considered).

2.2 Time Search Engine

A prototype of Web-search engine has been implemented using techniques of temporal information indexing through Lucene API (http://lucene.apache.org/) tools. The architecture is represented by Fig. 1. It is composed of two main modules (indexer and searcher), both relying on the TERN system.

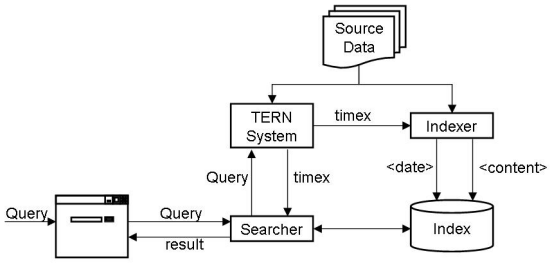

Fig. 1 Temporal information retrieval search engine architecture

The core of the indexer built with Lucene library takes all the terms from a set of source documents and filters them by Spanish stopwords to obtain the relevant ones. Afterwards, it creates an index with such terms that is increased with those temporal expressions normalized by the TERN system. The searcher module receives user queries in a textual format, through a Web interface. In a first step, such queries are processed by the TERN system. Once normalized, timexes together with the rest of

relevant terms of the query are used to launch the search to the index. From this search documents containing the query terms are retrieved and ordered applying relevance criteria defined by the scorer of Lucene.

3 System Architecture

A corpus with enough dates and temporal references is necessary to evaluate the benefits of a Web-search engine. We were not able to find a suitable collection so we decided to build a new corpus. Newswire from digital newspapers was considered a good domain, taking articles from sports news and cultural agenda due to the higher frequency of embedded temporal expressions. Selected sources are "www.elmundo.es" (72 documents) and "www.elpais.es" (45 documents). All news was dated in November, 2008. On the other hand, topics pointing the user information needs and relevance judgements must be defined. The first ones try to simulate a real information need related to the corpus subject. The second ones determine for each topic if a document is relevant or not. Topics have been classified according to 6 types of user potential information needs: theatre, cinema, music, football, tennis and basketball. Moreover, a topic must be restricted in time (e.g.: "tennis tomorrow").

The evaluation of the system has been carried out by comparing the performance of two search engines. The first one (baseline system) indexes relevant terms of the collection and looks for the exactly match with the query terms of the topics. The second has been improved with temporal information management capabilities thanks to the integration of the TERN system. Thus, both the query terms and the indexes are enriched with normalized temporal information, capturing the semantic of time.

In Table 1 a sample of the topics together with information about precision (P), TOP@5 values, total number of retrieved (# RETR) and relevant (# REL) documents for each topic is presented.

Table 1 Baseline vs. Temporal information search engine performance sample.

TOPIC		# REL	BASELINE			TEMPORAL		
EVENT	**TIME RESTR.**		**P(%)**	**top@5**	**#RETR**	**P (%)**	**top@5**	**# RETR**
teatro (*theatre*)	hoy (*today)*	3	5,7	0,2	53	11.5	0,4	26
	ayer (*yesterday)*	4	10,8	0	37	14.8	0,6	27
	este sábado(*this Saturday)*	3	9,1	0,4	33	10.7	0,4	28
fútbol (*soccer*)	hoy (*today)*	3	8,8	0,2	44	10.3	0,6	29
	ayer (*yesterday)*	3	8,3	0,2	36	9.7	0,6	31
	mañana (*tomorrow)*	4	12,1	0,4	33	13.3	0,6	30

As it can be observed, evaluated information needs allow a better performance to the temporal search engine in comparison to the baseline. Results show that more accurate hits are obtained using the temporal information search engine. In addition, top hits achieve a higher relevance score. With temporal management the search engine is able to detect that in some documents "*this Saturday"* is referred to the 21st

of November, in spite of it is not explicitly said in the contents. This is possible thanks to the analysis of the context and the resolution of their references.

As and additional result, the Mean Reciprocal Rank (MRR) for the baseline system is 0.40, whereas the temporal system is 0.81. An increase of MRR = 50%, is obtained in this evaluation.

4 Conclusions and Future Work

This work presents a Web-search engine that analyses the documents in order to extract temporal information and its semantics. This information is used to refine the results of the user queries. Searches run over the evaluation corpus show improved results, since the use of the semantics captured from the temporal information of documents makes the engine more powerful than in the case of using only their textual representation. In accordance with the results of this work, although preliminary, it can be said that temporal information management constitutes a line of improvement to be taken into account.

Several improvements should be made to this proposal. A larger corpus and more realistic queries are needed. Management of events has to be added to the TERN system, whose coverage of temporal expressions could also be enhanced. It is also need to research on context extraction mechanisms that facilitate the detection of temporal events and its resolution. Finally, it would be interesting to promote a common framework as testbed, together with the development of resources like specific corpora for evaluating the temporal information retrieval benefits.

Acknowledgments.
This work has been partially supported by the Research Network MAVIR (S-0505/TIC-0267), and project BRAVO (TIN2007-67407-C03-01).

5 References

1. Alonso, O. Gertz, M. and Baeza-Yates, R.: On the value of temporal information in information retrieval. ACM SIGIR Forum, vol. 41, no. 2, pp. 35--41, (2007).
2. Roddick, J. F. and Patrick, J. D.: Temporal semantics in information systems: a survey. Information Systems, vol. 17, no. 3 pp. 249--267, (1992).
3. Mani, I.and Wilson, G.: Robust Temporal Processing of News. In Proceedings of the ACL'2000 Conference. Hong Kong (2000).
4. Vicente-Díez, M.T., de Pablo-Sánchez, C. and Martínez, P.: Evaluación de un Sistema de Reconocimiento y Normalización de Expresiones Temporales en Español. In Proceedings of SEPLN 2007, pp. 113--120. Sevilla (2007).
5. Vicente-Díez, M.T., Samy, D. and Martínez, P.: An Empirical Approach to a Preliminary Successful Identification and Resolution of Temporal Expressions in Spanish News Corpora. In Proceedings of LREC'08. Marrakech (2008).

Uso de PLN en otras disciplinas

Gemma Boleda

Universidad Politécnica de Cataluña

> "Human knowledge is expressed in language. *So computationallinguistics is very important.*" (Mark Steedman, 2008, On becoming a discipline, *Computational Linguistics*, 34(1), p. 144)

Propongo utilizar herramientas, recursos y técnicas de PLN (*taggers*, *parsers*, recursos como WordNet, técnicas de aprendizaje automático) para la investigación en otras disciplinas. Como una gran parte del conocimiento humano sobre ámbitos muydiversos está codificado en textos escritos, es útil procesar dichos textos con herramientas que permiten por ejemplo lematizar o identificar las palabras más importantes de un texto. Recientemente se ha iniciado esta vía en campos como el denominado *Biomedical NLP* (Ananiadou y McNaught 2006); propongo extender este tipo de hibridaciones a más campos.

La aplicación más inmediata, por el objeto de estudio, es en Lingüística, aunque resulta sorprendente la poca interacción entre PLN y Lingüística teórica (Sparck-Jones 2007). Respecto a este campo, resumiré la metodología y los resultados de dos proyectos en los que he participado (categorización semántica de adjetivos y estudio de las construcciones con expresiones de nacionalidad). Sin embargo, las posibilidades son amplias y variadas, y abarcan desde las ciencias sociales hasta la física. Ejemplificaré dichas posibilidades con una colaboración en curso entre físicos, ingenieros y lingüistas en que estamos estudiando la distribución de las distancias entre palabras como sistema complejo, y encontramos similitudes sorprendentes entre dicha distribución y la de fenómenos catastróficos como los terremotos.

Los puntos a favor de este tipo de hibridaciones son tanto científicos, pues se promociona la interdisciplinariedad "de verdad", como prácticos, pues permiten a los científicos publicar en revistas de alto impacto que les serían si no vedadas y a las instituciones recibir fondos de más fuentes. Con este tipo de colaboraciones se promociona también la visibilidad social del PLN, y a medio plazo sus aplicaciones industriales.

Sin embargo, hay experiencias recientes que muestran las dificultades de la integración del PLN en otros ámbitos, como la Recuperación de Información (IR): inicialmente, se mostró que procesos como la lematización o la categorización no mejoraban o incluso empeoraban los resultados de IR. Sin embargo, en investigaciones y desarrollos recientes los resultados son más prometedores (BaezaYates 2008). Para poderse usar como herramientas en otros ámbitos, las herramientas y técnicas de PLN deben ser más robustas y completas, así como aumentar su precisión.

REFERENCIAS

Ananiadou, Sophia y John McNaught (editores). 2006. *Text Mining for Biology and Biomedicine*. Boston and London: Artech Hose.

BaezaYates, Ricardo. 2008. From Capturing Semantics to Semantic Search: A Virtuous Cycle. *The Semantic Web: Research and Applications*. *Lecture Notes in Computer Science*. Berlin / Heidelberg: Springer.

Spärck Jones , Karen. 2007. Computational Linguistics: What About the Linguistics? *Computational Linguistics*, Volume 33, pp. 437441.

Búsqueda y navegación

Expansión de Consultas, basado en PRF

Joaquín Pérez Iglesias
NLP&IR Group, UNED
joaquin.perez@lsi.uned.es

Lourdes Araujo Serna
NLP&IR Group, UNED
lurdes@lsi.uned.es

José R. Pérez-Agüera
Universidad Complutense de Madrid
jose.aguera@fdi.ucm.es

12 de enero de 2009

La investigación se centra en expansión de consultas, más específicamente en "pseudo-relevance feedback". La expansión de consultas basada en PRF, consiste en extraer el conjunto de términos expandidos de manera automática a partir del subconjunto de documentos más relevantes recuperados con la consulta original. Finalmente se realiza una nueva consulta formada por los términos originales junto a los términos expandidos.

En la literatura relacionada se observa que la aplicación de esta técnica produce, en general [CdMRB01], una mejora en la calidad de los resultados especialmente en términos de cobertura, pero con un análisis más en detalle de los resultados obtenidos se observa que para ciertas consultas la aplicación de esta técnica degrada la precisión de los resultados.

Nuestra línea principal de investigación consiste en analizar en profundidad que hace que para ciertas consultas la expansión degrade la calidad de los resultados, y a partir de esta información formular un método de expansión que permita mejorar las técnicas utilizadas actualmente.

Un primer paso necesario para llevar a cabo el análisis propuesto consiste en la detección de aquellos términos expandidos que junto a la consulta original nos permitan obtener los valores de MAP más elevados posibles. Para la obtención de dichos términos se plantea la utilización de un algoritmo genético, que calcule aquella combinación de términos expandidos que maximize el valor de MAP, para un consulta específica.

Una vez obtenidos aquellos términos de expansión óptimos, el objetivo se centra en detectar que características hacen que una combinación de términos expecífica sea la ideal para realizar la expansión. Las características que utilizamos son de tipo estadístico, es por ello que los datos de observación que se utilizan, son: frecuencia de términos, frecuencia inversa, frecuencia en colección, análisis de coocurrencia entre los términos, etc... Una consideración clave que surge a partir de este enfoque es hasta qué punto y utilizando como base el mero análisis estadístico será posible resolver este problema, o si por el contrario esta información no será suficiente para abarcar el problema planteado.

Una primera aproximación para detectar aquellos rasgos de caracter estadístico que puedan definir la calidad de un conjunto de términos de expansión podría estar basada en la construcción de un clasificador, utilizando como datos

de entrenamiento aquellos conjuntos de términos que maximizan el MAP.

Un enfoque similar al descrito se puede encontrar en [CNGR08], donde se describe la construcción de un clasificador, basado en los rasgos estadísticos de los términos. Este clasificador se entrena partiendo de tres subconjuntos: términos óptimos para la expansión, términos que no influyen en la expansión y finalmente aquellos términos negativos que introducen ruido en el proceso de recuperación. El enfoque planteado en nuestra investigación, con el uso de algoritmos genéticos, aporta una clara mejora al descrito por Robertson en el que los términos son tratados de forma independiente entre sí, obviando explícitamente el hecho de que en general son las combinaciones entre los términos, y no estos de forma individual, los que provocan la mejora o degradación en la calidad de los resultados.

Es de esperar que como resultado de la construcción del clasificador así como de la realización de un análisis pormenorizado de los datos obtenidos, se pueda dar una primera respuesta a algunas de las dudas que surgen con la aplicación de la técnica de PRF en la expansión de consultas.

Referencias

[CdMRB01] Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, January 2001.

[CNGR08] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *SIGIR*, pages 243–250. ACM, 2008.

Content-based Clustering for Tag Cloud Visualization

Arkaitz Zubiaga, Alberto P. García-Plaza, Víctor Fresno y Raquel Martínez

ETSI Informática
Universidad Nacional de Educación a Distancia

Social tagging systems are becoming an interesting way to retrieve previously annotated data. The huge amount of users tagging web content creates useful metadata for it, facilitating the subsequent access to it. Nevertheless, the non-hierarchical nature of its classification scheme requires an additional organization of the tags in order to correctly navigate between them. This type of sites present, as a way to navigate, a tag cloud made up by the most popular tags on its collection, where nor tag grouping neither their corresponding content is considered.

In this paper, we present a methodology to obtain and visualize a cloud of related tags based on the use of self-organizing maps. Unlike other previous works, our methodology establishes the relations among tags based on the textual content of the annotated web documents, with no information about tag co-occurrence. Each unit of the map can be represented by the most relevant terms of the tags it contains, so that it is possible to study and analyze the groups as well as to visualize and navigate through the relevant terms and tags.

Finally, we make a qualitative evaluation of the obtained results, and present clustered tag cloud to ease the navigation. In addition, we introduce DeliciousT140, a dataset we generated for these experiments, based on the 140 most popular tags offered by the social bookmarking site Delicious.

Combinación de técnicas textuales y visuales para la recuperación de imágenes

Rubén Granados Muñoz (1) y Ana García Serrano (2)

(1) Facultad de Informática, Universidad Politécnica de Madrid
(2) ETSI Informática, Universidad Nacional de Educación a Distancia
rgranados@fi.upm.es, agarcia@lsi.uned.es

En el grupo GSI-ISYS[i] se ha desarrollado una herramienta de indexación y recuperación de información siguiendo una estructura modular, que permite analizar y evaluar el comportamiento del sistema ante cualquier adaptación/ampliación/cambio sobre el esquema base. La implementación de esta herramienta, denominada IDRA, ha ido evolucionando hasta realizarse íntegramente en Java. La indexación y búsqueda de documentos que sirve como baseline, se realiza siguiendo el modelo clásico de espacio vectorial con una función tf-idf normalizada, sobre diferentes formatos de colecciones de documentos.

Durante el proceso de construcción de esta herramienta, y como parte de los fines propuestos en el proyecto coordinado BRAVO (TIN2007-67407-C03-03), se persigue la flexibilidad para incorporar y variar diferentes parámetros, así como la optimización de la herramienta IDRA, en cuanto al tiempo y espacio necesarios para los procesos de indexación, y las medidas de recuperación. También se puede mencionar la inclusión de ontologías de dominio y entidades nombradas en la recuperación orientada a dominios específicos.

Actualmente, la investigación se está focalizando en temas relacionados con la recuperación de información multimedia y, más concretamente, en la Recuperación de Imágenes para analizar las posibilidades de mejora de la colaboración entre técnicas de recuperación textual y visual. Para esto, se ha formado un grupo de trabajo con investigadores de la Universidad de Valencia, que tienen desarrollado un sistema CBIR (Content Based Image Retrieval), con los que ya se ha trabajado anteriormente y que también pueden variar los métodos basados en características visuales.

Se ha participado en el foro de evaluación competitiva CLEF 2009, en la tarea ImageCLEF de recuperación fotográfica (ImageCLEFphoto) con el objetivo tanto de evaluar el comportamiento de la recuperación textual de la herramienta IDRA utilizando las anotaciones de las imágenes de la colección, como de la combinación con distintas configuraciones del sistema CBIR de la Universidad de Valencia. Los resultados obtenidos mostraban un comportamiento mucho mejor del experimento basado únicamente en texto que todos aquellos que utilizaban sólo la recuperación visual. En cambio, algunos de los experimentos realizados con las distintas técnicas de fusión de resultados, mejoraban los obtenidos por el experimento textual base. Este año, con la

mejora de los sistemas individualmente, y nuevas formas de combinación se espera poder mejorar significativamente los resultados obtenidos.

Además, se pretende avanzar en el estado del arte en temas relacionados con la anotación automática de contenido multimedia (AIA, Automatic Image Annotation). En este campo, se está intentando definir un conjunto de conceptos semánticos visuales para poder etiquetar o clasificar imágenes. Estos conceptos pueden presentarse en forma de simples vocabularios controlados o, jerárquicamente (como por ejemplo LSCOM para multimedia). Habitualmente, una vez definidos estos conceptos, y mediante técnicas de aprendizaje automático (normalmente SVM, máquinas de vectores de soporte), se construyen detectores automáticos para cada uno de ellos (por ejemplo: Columbia374 o Mediamill). Es la información proporcionada por estos detectores la que pretende ser utilizada para mejorar la herramienta IDRA y para evaluar dichos resultados se participará en varias tareas del ImageCLEF 2009.

[i] El grupo de investigación GSI-ISYS (Grupo de Sistemas Inteligentes – Intelligent SYStems), es un grupo consolidado de la Universidad Politécnica de Madrid.

Towards the Evaluation of Voice-Activated Question Answering Systems: Spontaneous Questions for QAST 2009

Davide Buscaldi[1] and Paolo Rosso[1] and Jordi Turmo[2] and Pere R. Comas[2]

[1] Natural Language Engineering Lab., RFIA
Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain,
{dbuscaldi,prosso}@dsic.upv.es
[2] TALP Research Center,
Universitat Politècnica de Catalunya
{turmo,pcomas}@lsi.upc.edu

Abstract. This is a preliminary report of the work carried out in order to introduce "spontaneous" questions into QAST at CLEF 2009. QAST (Question Answering in Speech Transcripts) is a track of the CLEF campaign. The aim of this report is to show how difficult can be to generate "spontaneous" questions and the importance to take into account the real information needs of users for the evaluation of question answering systems.

1 Introduction

In the Question Answering (QA) task, search engines have to extract concise and precise fragments of texts that contain an answer to a question posed by the user in natural language. This task is very close to what is usually considered as "automatic text understanding".

The availability of effective QA systems may change the type of interaction between humans and machines, mainly thanks to the fact that in QA the user obtains an answer and not a list of documents to be browsed. Competitions like CLEF (http://www.clef-campaign.org/) have been created in order to develop and improve existing systems and to evaluate and compare their behavior.

A potentially interesting evolution of QA systems is their application to spoken documents. Until now, most QA research has focused on mining document collections containing written texts to answer written questions [2]. They usually share a decent writing quality, at least grammar-wise. In addition to these written sources, a lot of potentially interesting information appears in spoken documents, such as broadcast news, speeches, seminars, meetings or telephone conversations. The QAST track has been introduced with the objective of investigating the problem of question answering in such audio documents [4].

A major step towards a more natural type of interaction between machines and users will be represented by the introduction of Speech Language Technologies (SLT) into QA systems [1,3]. Automatic Speech Recognition (ASR) and the

development of spoken human-machine interfaces are currently considered mature enough to be used in most common applications. Some examples of these systems are represented by spoken QA systems that could be used by means of mobile devices.

The realization of this kind of systems will meet with a series of issues, due to the nature of the input medium, that is particularly sensitive to errors. We propose to introduce spontaneous oral questions into QAST 2009. In most QA evaluations, questions have been posed in good quality, without errors and with a clear definition of focus and topic. Voice-activated systems will have to face with questions that include misspelled names, pauses, hesitations.

Unfortunately, until now, a corpus of questions of this kind has not been released. This is not only due to the usual problems for the production of corpora, especially for the time necessary to collect the questions or the money needed to develop the resource. In this report we will give an overview of the issues of producing spontaneous oral questions.

2 QAST 2009 proposal

The objective of this pilot track is to develop a framework in which QA systems can be evaluated when the answers have to be found in speech transcripts. There are three main objectives to this evaluation:

1. Motivating and driving the design of novel and robust QA architectures for speech transcripts and voice-activated systems;
2. Measuring the loss due to the inaccuracies in state-of-the-art ASR technology;
3. Measuring this loss at different ASR performance levels given by the ASR word error rate;

The collections will be composed by the 2005 and 2006 TC-STAR European Parliament Plenary Sessions (EPPS) corpora, in English and Spanish, and the ESTER French broadcast news corpus, with automatic and manual transcriptions.

The types of definitional questions for the task will be 4: *Person*, *Organisation*, *Object* and *Other*. There are 10 types of factual questions: *Person*, *Location*, *Organisation*, *Time* (including dates), *Measure*, *System*, *Language*, *Colour*, *Shape* and *Material*. Questions will be both "standard" written questions and "spontaneous" oral questions, manually transcribed.

The organisation of the track will be carried out by the UPC (Universitat Politècnica de Catalunya), together with the LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur) and the UPV (Universidad Politécnica de Valencia).

3 Issues in the production of spontaneous questions

Two tests were carried out with 4 Spanish speakers, who were requested to pick one or more text fragments from the EPPS transcriptions of days 15-18

Nov. 2004 and formulate two or more questions over each text. In the first test they were instructed to ask for something that they felt it was missing for the understanding of the text. In the second test they were requested to formulate their questions following the task guidelines.

In the first test, the users produced 17 questions, an average of 4.25 per user, using 6 text fragments (2.83 questions per text fragment), employing an average of 1 : 15 minutes for each questions.

In Table 1 we show a sample of the produced questions.

Table 1. First test: completely free questions.

Question
3. Quién es Annetta Flanigan?
4. Permiten a los rehenes establecer comunicacin con el exterior?
6. Porqué acusan de omicidio aaa ... Dow Chemical?
7. Cuánta gente ha muerto ... a causa de los sucesos relacionados con esta empresa?
8. Er... quién les ha hecho el embargo a China?
12. Quién lee este texto?
15. Cuál es la política del parlamento europeo sobre ... el cambio climático?
17. A quién va dirigido el texto?

As it can be observed from this sample, most questions do not respect guidelines, contain anaphoras (4 and 7), ask about the speaker or the audience (12 and 17) or contain hesitations, pauses and misspellings (6,7,8 and 15).

In the second test, the users were able to produce questions that better fit the guidelines, although in many cases there were still anaphoras (see Table 2, questions 10 and 15) and "meta"-questions (i.e., questions about the speech and not the content, such as question 5). The main problem, in this case, is due to the fact that most questions did not have an answer in the collection.

Table 2. Second test: users instructed with guidelines.

Question
1. Cómo se llama el presidente de China?
5. En qué fecha se sitúa el texto?
6. Quién es el presidente de la delegación de Estados Unidos?
9. Quién es el señor Buttiglione?
10. Cuál es su postura?
14. Quién es ... el nombre del defensor del pueblo?
15. Hay uno en el parlamento europeo para toda Europa?

4 Conclusions

The evaluation showed that it is possible to produce a set of questions with the methodology proposed, even if three main problems emerged:

1. Users do not stick to the guidelines even if they were told to, with the result that a number of questions cannot be used;
2. The formulated questions contain a high percentage of NIL questions;
3. The production of the text fragments could be very demanding if it is not possible to guide users to produce more "good" questions.

Acknowledgements

The introduction of spontaneous questions in QAST 2009 is the result of the collaboration between UPV and UPC in the context of the TextMESS research project (TIN2006-15265-C06).

References

1. Dan Moldovan. Voice activated question answering. In *IEEE Workshop on Spoken Language Technology*, pages 5–5, Palm Beach, CA, 2006.
2. Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors. *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers*, volume 4730 of *Lecture Notes in Computer Science*. Springer, 2007.
3. Emilio Sanchis, Davide Buscaldi, Sergio Grau, Lluis Hurtado, and David Griol. Spoken qa based on a passage retrieval engine. In *IEEE-ACL Workshop on Spoken Language Technology*, pages 62–65, Aruba, 2006.
4. Jordi Turmo, Pere R. Comas, Sophie Rosset, Lori Lamel, Nicolas Moreau, and Djamel Mostefa. Overview of qast 2008. In *CLEF 2008 Working Notes*, Aarhus, Denmark, 2008. TrebleCLEF.

Combinación de técnicas lingüísticas y estadísticas para la generación de resúmenes automáticos*

Elena Lloret y Manuel Palomar

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{elloret, mpalomar}@dlsi.ua.es

1. Introducción

Este trabajo contiene un resumen de la investigación que se está desarollando y, en concreto, este artículo se centra en la producción de resúmenes monodocumento de tipo extractivo para documentos en inglés pertenecientes al dominio periodístico. El objetivo es presentar un nuevo método (con dos variantes) para determinar la relevancia de las frases de un documento, con la finalidad de decidir si deben o no formar parte del resumen final. La hipótesis de este trabajo es que la combinación de técnicas basadas en teorías lingüísticas (el *Principio de la Cantidad de Codificación* (CQP[1])) [1] junto con técnicas estadísticas (concretamente, la frecuencia de las palabras) da lugar a que los resúmenes generados contengan información relevante, evitando así que incorporen información ruidosa o carente de relevancia.

El trabajo se estructura de la siguiente manera: en la sección 2 describimos las dos aproximaciones del nuevo método que proponemos, basado en el *CQP*, para detectar las frases más importantes en un documento. Posteriomente, la sección 3 contiene los experimentos realizados y los resultados obtenidos. Finalmente, las principales conclusiones obtenidas y los trabajos futuros que se plantean se exponen en la sección 4.

2. El Principio de la Cantidad de Codificación

El ***Pincipio de la Cantidad de Codificación*** *(CQP)* descrito en [1] es una teoría de origen lingüístico que sostiene que la información codificada con más unidades resalta sobre toda la demás, haciendo que el lector centre su atención sobre dicha información, y por tanto, haciendo que ésta se recuerde posteriormente con mayor facilidad. Esto se puede interpretar como que la información más importante de un texto se expresará con mayor cantidad de elementos (por ejemplo, sílabas, palabras o sintagmas). En [2] se ha demostrado que este principio se cumple en textos escritos. Por otro lado, un sintagma nominal es el tipo de estructura sintáctica capaz de expresar más o menos información en función del número de palabras que contenga (ya que puede estar formado por pronombres, nombres, adjetivos e incluso relaciones de relativo).

* Esta investigación ha sido financiada por el Ministerio de Ciencia e Innovación bajo el proyecto TEXT-MESS (TIN2006-15265-C06-01).

[1] De las siglas en inglés de **C**ode **Q**uantity **P**rinciple.

A partir de estas afirmaciones, el enfoque que se propone en este trabajo es estudiar el grado en el que la utilización de este principio puede ser adecuado en la tarea de generación automática de resúmenes. Para ello, decidiremos qué frases son las más relevantes de un texto considerando la cantidad de información léxica que incluyen, proponiendo dos variantes para este enfoque. En primer lugar, vamos a estudiar el *CQP* de la manera más simple (es decir, teniendo en cuenta solamente la codificación empleada en cada sintagma nominal), mientras que en una segunda variante, consideraremos además la frecuencia de las palabras que integran dichos sintagmas nominales, de tal manera que tendrán más peso aquellos sintagmas nominales que contengan palabras más frecuentes. El uso de la frecuencia de las palabras queda justificado en [3] y [4], donde se demuestra que esta técnica influye positivamente en la producción de resúmenes automáticos.

La hipótesis que vamos a investigar en este trabajo es la siguiente: las frases que contengan sintagmas nominales más largos (variante 1) o con palabras de mayor frecuencia (variante 2), recibirán mayor puntuación, dando lugar a que las oraciones que más puntuación tengan en un documento sean seleccionadas para formar parte del resumen final. Para identificar los sintagmas nominales usamos la herramienta *BaseNP Chunker* desarrollada en la Universidad de Pennsylvania. Adicionalmente, para esta investigación consideramos las palabras (sin tener en cuenta las palabras vacías) como unidades léxicas. La fórmula 1 muestra la forma de puntuar cada oración de un documento de acuerdo a la variante 1. La puntuación de una frase aumentará en una unidad cada vez que una palabra pertenezca a un sintagma nominal y se normalizará respecto al número total de sintagmas nominales que contenga.

$$Sc_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |w| \ . \tag{1}$$

La fórmula 2 representa la variante 2, donde se tiene en cuenta también la frecuencia de las palabras para puntuar las frases (ahora no se suma una unidad por cada palabra perteneciente al sintagma monimal, sino que el peso de cada palabra será la frecuencia de la misma en el documento).

$$Sc_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} tf_w \ . \tag{2}$$

3. Experimentación y resultados

Evaluar un resumen, ya sea de forma manual o automática, es una tarea costosa y difícil puesto que un resumen es subjetivo y dependiendo de las necesidades del usuario será más o menos útil. La principal dificultad a la hora de la evaluación proviene de la imposibilidad de construir un resumen que constituya modelo único y adecuado con el que poder comparar un resumen automático [5]. Además, podría darse el caso de que un resumen automático fuera correcto aunque no se asemejara a ningún resumen modelo generado [6].

Para la investigación llevada a cabo en este trabajo realizamos la evaluación con la herramienta ROUGE [7], que ha sido muy utilizada para medir automáticamente el grado de solapamiento entre el vocabulario de un resumen producido por un humano y uno generado de forma automática, atendiendo a diversos criterios (unigramas, bigramas, longitud de la cadena más larga, etc.). Tomando como punto de partida los 567 documentos pertenecientes al corpus del DUC 2002, generamos los resúmenes automáticos siguiendo las directrices establecidas para una de las tareas de dicha competición, y aplicando las variantes del método presentado en la sección 2. Para poder comparar nuestros resultados con los obtenidos por los participantes de ese año, tomamos como referencia los resultados recogidos en [8], donde se volvieron a evaluar los sistemas utilizando la herramienta ROUGE.

La tabla 3 muestra los resultados (*recall*) de los dos mejores sistemas del DUC 2002 (S28, S21), del baseline propuesto[2] y de nuestras aproximaciones (*CQPSum* y *CQPSumTf*, para las variantes 1 y 2, respectivamente) evaluados sobre ROUGE-1 (unigramas), ROUGE-2 (bigramas), ROUGE-SU4 (bigramas no consecutivos) y ROUGE-L (subsecuencia común más larga). Como se observa en la tabla, el sistema S28 obtuvo los mejores resultados en el DUC 2002. Los resultados obtenidos para nuestra primera aproximación son ligeramente inferiores que para el mejor sistema del DUC 2002. Sin embargo, cuando ejecutamos la segunda variante de nuestro método, se puede observar como estos resultados mejoran considerablemente para todos los valores de ROUGE, excepto para ROUGE-2. Si analizamos los incrementos de mejora obtenidos, observamos que cuando combinamos la técnica de frecuencia de palabras con el *CQP*, el incremento de mejora obtenido es de un 4 % sobre el mejor sistema del DUC (S28), mientras que mejora obtenida respecto a nuestra primera variante es de un 7,23 %. De esta manera se observa que la combinación de técnicas basadas en principios linüísticos con otras técnicas, como pueden ser técnicas estadísticas, es adecuado para la tarea de generación automática de resúmenes.

Cuadro 1. Resultados para las aproximaciones propuestas.

SYSTEM	ROUGE-1	ROUGE-2	ROUGE-SU4	ROUGE-L
CQPSumTf	**0,44565**	0,18738	**0,20993**	**0,40295**
S28	0,42776	0,21769	0,17315	0,38645
CQPSum	0,42241	0,17177	0,19320	0,38133
S21	0,41488	0,21038	0,16546	0,37543
DUC baseline	0,41132	0,21075	0,16604	0,37535

[2] El baseline del DUC consistía en formar el resumen seleccionando las primeras 100 palabras del documento.

4. Conclusiones y trabajos futuros

El objetivo de este trabajo era presentar un nuevo método para calcular la relevancia de las oraciones en un documento. El método presentado se basa en el *Principio de la Cantidad de Codificación* que analiza cómo los humanos codifican la información de una u otra manera al escribir, según las ideas que quieran resaltar más. En esta investigación se ha llevado a cabo una sencilla implementación de esta teoría lingüística combinada con una técnica estadística (cálculo de la frecuencia de las palabras) y se ha ejecutado sobre el corpus de datos del DUC 2002, obteniendo resultados prometedores al evaluar los resúmenes con la herramienta ROUGE.

Como trabajos futuros, se plantea adaptar este método a la generación de resúmenes multidocumento, además de estudiar cómo se puede aportar más conocimiento a esta nueva técnica con la finalidad de mejorar la calidad de los resúmenes automáticos en cuanto a información contenida se refiere, con el objetivo final de construir un sistema de generación automático de resúmenes completo y robusto. Otro aspecto muy importante a tener en cuenta de cara a posibles trabajos futuros es diseñar y estudiar diversos criterios que son imprescindibles tener en cuenta a la hora de evaluar la calidad de un resumen, como por ejemplo, la coherencia de un texto, gramaticalidad o redundancia, y no centrarnos solamente en evaluaciones cuantitativas del contenido del resumen.

Referencias

1. Givón, T.: A functional-typological introduction, II. Amsterdam : John Benjamins (1990)
2. Ji, S.: A textual perspective on Givon's quantity principle. Journal of Pragmatics **39**(2) (2007) 292–304
3. Lloret, E., Ferrández, O., Muñoz, R., Palomar, M.: Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) (41) (2008) 183–190
4. Nenkova, A., Vanderwende, L., McKeown, K.: A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In: SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. (2006) 573–580
5. Fuentes Fort, M.: A Flexible Multitask Summarizer for Documents from Different Media, Domain, and Language. PhD thesis (2008) Adviser-Horacio Rodríguez.
6. Mani, I.: Summarization evaluation: An overview. In: Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL). Workshop on Automatic Summarization. (2001)
7. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003). (2003) 71–78
8. Steinberger, J., Poesio, M., Kabadjov, M.A., Ježek, K.: Two uses of anaphora resolution in summarization. Information Processing & Management **43**(6) (2007) 1663–1680

Detecting Drug-Target articles by NLP techniques[1]

Roxana Danger[1], Isabel Segura-Bedmar[2], Paloma Martínez[2], Paolo Rosso[1]

[1] Universidad Politécnica de Valencia
[2] Universidad Carlos III de Madrid
{rdanger,prosso}@dsic.upv.es}{isegura,pmf}@inf.uc3m.es}

Abstract.

Important progress in treating diseases like cancer, AIDS or Parkinson's disease has been possible thanks to the identification of drug targets. However, the knowledge about drugs, their mechanisms of action and drug targets is hidden in the millions of biomedical articles. This paper describes an approach for text classification which combines text processing, semantic information from biomedical resources and machine learning techniques to identify drug targets articles, which is the first issue in order to make all drug target information available for medical researchers.

Keywords: Biomedical Text Classification, Machine Learning, UMLS, MetaMap, Drug Targets.

1. Introduction

A drug target is a defined molecule or structure within the organism, that is linked to a particular disease, and whose activity can be modified by a drug. For disease intervention, the drug target may be either activated or inhibited by a drug. The approaches used for the design of drugs, firstly, study the disease process and determine its physiologic mechanics to detect the drug targets related to this disease. Then, new drugs are designed to act on these targets.

In recent years, important progress in treating diseases like cancer, AIDS or Parkinson's disease has been possible thanks to the identification of drug targets for these diseases [1], [6], [3]. With long and costly drug development times there is a need in the pharmaceutical industry to prioritize targets early in the drug discovery process.

In the last two decades, our knowledge about drugs, their mechanisms of action and drug targets has rapidly increased. Nevertheless, this knowledge is far from being complete and is highly fragmented because most of it is hidden in millions of medical articles and textbooks. Extracting knowledge from this large amount of unstructured information is a laborious job, even for human experts.

Motivated by this fact, we have developed a novel approach for text classification which combines text processing, semantic information from biomedical resources and

[1] This research work is supported by projects TIN2007-67407-C03-01 and S-0505/TIC-0267, as well as for the Juan de la Cierva program of the MICINN of Spain.

machine learning techniques to identify drug targets articles. Currently, we are evaluating the classification capability of several supervised algorithms such as Support Vector Machines (SVM) or K nearest neighbor (KNN) as methods for locating articles about drug targets.

We think that our work could assist in constructing more complete and relevant drug targets databases, and so, the pharmaceutical industry could profit by exploiting these resources.

2. Proposal

The first problem to solve for the drug target article classification problem is to construct a corpus, as it is not available. We have built a corpus of "positive" and "negative" drug targets articles (137 positives and 2864 negatives) from DrugBank[2] and Pubmed.

DrugBank is an annotated database that combines detailed chemical, pharmaceutical and pharmacology drug and target information. DrugBank contains 4900 drug entries. Each entry contains more than 100 data fields with chemical, pharmacological, pharmacogenomic and molecular biological information. In particular, 14000 drug targets are linked to these drug entries. Pubmed[3] is a bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. It contains about 18 million references to journal articles in life sciences. Pubmed database contains efficient tools for searching and annotating documents. Each article is described by information such as authors, journal of the publication, year, etc., and even more important: the authors of an article can associate to it a set of UMLS[4] (Unified Medical Language System) terms which characterize it. UMLS is a set of resources developed by the National Library of Medicine (NLM) whose main objective is to assist in the developing of natural language technology for biomedical texts. UMLS Methathesaurus contains a large controlled vocabulary related to all biomedical themes, and maintains links to terms described in others biomedical dictionaries.

DrugBank database contains articles from 1995 to 2003, with a highest frequency between 1995 and 2001. Therefore, the corpus was created with articles in this range of years. About a 5% of all articles in Pubmed are concerning to drug targets. Such distribution was measured querying Pubmed about articles with the UMLS synonyms of the term "biological target". In this way, an article was marked as a drug target one if it contains (or was annotated in Pubmed with) at least one of these synonyms. Using the same distribution which was observed all Pubmed, a corpus of 3001 article has been created, The 137 positive articles were randomly selected from the articles of DrugBank, and the 2864 negative articles were randomly selected between articles of Pubmed, which were not marked as drug target article. Both sets contain only articles in the range from 1995 to 2001, and the distribution amongst drug target and no drug target articles observed in Pubmed for each year was maintained.

---

[2] http://www.drugbank.ca

[3] http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed

[4] http://www.nlm.nih.gov/research/umls/.

Each abstract was processed by the DrugNer [5] for drug name recognition and classification in biomedical texts. DrugNer uses the MetaMap Transfer (MMTx) [2], a program for syntactically parsing the text and linking the phrases that occur in the documents with concepts and semantic types defined by the Unified Medical Language System (UMLS). In addition, DrugNer also implements a set of nomenclature rules recommended by the World Health Organization (WHO) International Nonproprietary Names (INNs)[5] Program to identify and classify pharmaceutical substances that occur in the text. In particular, these rules consist of a set of affixes that are matched with each phrase in order to detect possible drugs and their pharmacological families.

The set of collected features used to construct the final dataset for drug target article classification are summarized as follows:

1. *Chemical terms*: UMLS terms about drugs and chemical products used by the authors to characterize their article (extracted from the field MESH of Pubmed database),
2. *Mesh terms*: other UMLS terms, different from the chemical terms, used by the authors to characterize their article (extracted from the field MESH of Pubmed database),
3. *Stems Title*: the stemmed words of the title (extracted by using a Porter stemmer).
4. *Stems Abstract*: the stemmed words of the abtstract (extracted by using a Porter stemmer).
5. *Family Drug*: the pharmacological families of the drugs mentioned in the abstract (extracted by using DrugNer tool [5]),
6. *Semantic types that occur in the title*: semantic types of the UMLS terms mentioned in the title (extracted by using MMTx [2]),
7. *Semantic types that occur in the abstract*: semantic types of the UMLS terms mentioned in the abstract (extracted by using MMTx [2]),

The first four features were converted as boolean data, describing whether a term appears in the respective attribute. The remainder features were converted as integer data, describing the frequency with which a term appears in the respective attribute. These transformations generate vectors of 32049 attributes. Finally, we obtain a database of 3001 articles classified as drug target or not drug target article and characterized by 32049 attributes.

3. Experimental results

An attribute selection process has to be done in order to reduce the high number of attributes. Particularly, we have use: CfsSubsetEval and FilteredSubsetEval algorithms. The classification has been performed by using: NNge, IB1, J48 and Bayes Net algorithms. A summary of these results are given in Table 1.

The preliminary results are quite good, especially considering the unbalanced distribution of the data, that is, 95% of the articles of database are representing negative examples.

5 http://www.who.int/medicines/services/inn/en/.

Table 1. Preliminary results of attribute selection and classification algorithms.

Classification Alg. (Attribute Selection Alg.)	F-measure
NNge(CfsSubsetEval, 40 attributes)	0.834
IB1(FilteredSubsetEval, 108attributes)	0.811
J48 (CfsSubsetEval)	0.775
Bayes Net (CfsSubset Eval)	0.774
SMO (CfsSubset Eval)	0.763

4. Conclusions and Future Work

This paper describes an approach for text classification which combines text processing, semantic information from biomedical resources and machine learning techniques to identify drug targets articles, which is the first issue in order to make all drug target information available for medical researchers. The preliminary results are quite good, especially considering that the unbalanced distribution of the data, that is, 95% of the database are representing negative examples. An exhaustive experimentation will be done in order to obtain an optimal solution based on the promising results here described.

5. References

1. Adam J. Adler. Mechanisms of T Cell Tolerance and Suppression in Cancer Mediated by Tumor-Associated Antigens and Hormones. Current Cancer Drug Targets. Vol 7, N 1, 3--14. (2007).
2. Aronson, A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc. AMIA Symp. 17, 17–21.
3. Bean, P., (2005). New drug targets for HIV. Clinical Infectious Diseases, 2005 - UChicago Press, Volume 41, Number S1, Pp.96-100
4. Imming,P., Sinning,C. and Meyer,A. (2006) Drugs, their targets and the nature and number of drug targets. Nat. Rev. Drug Discov., 5, 821–834.
5. Segura-Bedmar,I., Martínez,P., Segura-Bedmar,M.,. Drug name recognition and classification in biomedical texts A case study outlining approaches underpinning automated systems. Drug Discovery Today, Volume 13, Issues 17-18, September 2008, Pages 816-823
6. Vincenzo Di Matteo and Ennio Esposit (2003). Biochemical and Therapeutic Effects of Antioxidants in the Treatment of Alzheimer's Disease, Parkinson's Disease, and Amyotrophic Lateral Sclerosis. *Current Drug Targets-CNS & Neurological Disorders*, Volume 2, Number 2, Pp. 95-107

Interacción y ontologías para el acceso a servicios

AnHitz, development and integration of language, speech and visual technologies for Basque

VICOMTech, Elhuyar Foundation, Robotiker,

Aholab Group and IXA Group - University of the Basque Country

AnHitz is a project promoted by the Basque Government to develop language technologies for the Basque language. AnHitz is a collaborative project between five participants (VICOMTech (www.vicomtech.org), Elhuyar Foundation (www.elhuyar.org), Robotiker (www.robotiker.com), the IXA Group (ixa.si.ehu.es), the Aholab Signal Processing Laboratory Group (aholab.ehu.es)) with very different backgrounds: text processing, speech processing and multimedia. The project aims to further develop existing language, speech and visual technologies for Basque, but also, we are integrating such resources and tools into a content management application for Basque with a natural language communication interface.

This application consists of a Question Answering and a Cross Lingual Information Retrieval system on the area of Science and Technology. The interaction between the system and the user will be in Basque (the results of the CLIR module that are not in Basque will be translated through Machine Translation), using Speech Synthesis, Automatic Speech Recognition and a Visual Interface.

These are the features of the system we have built:

- The system will simulate an expert on Science and Technology. It will be able to answer questions or retrieve documents containing some search terms using a multilingual knowledge base.
- It will automatically translate the results to Basque if they are in English or Spanish.
- The interaction with it will be via speech. We will talk to it in Basque, and it will answer speaking in Basque too.
- The system will have a 3D human avatar that will show emotions depending on the success obtained in accomplishing the task.

The system use the following modules as can be seen in Fig. 1.:

- A 3D Human Avatar expressing emotions, developed by VICOMTech.
- A Basque Text-To-Speech synthesizer (TTS), developed by Aholab.
- Two Basque Automatic Speech Recognition systems (ASR), one of them developed by Aholab, the other one integrated by Robotiker.
- A Basque Question Answering system (QA), developed by IXA, over a Science and Technology knowledge base, compiled by Elhuyar.

- A Basque-Spanish-English Cross-Lingual Information Retrieval system (CLIR), developed by Elhuyar, over a Basque-Spanish-English comparable corpus on Science and Technology, compiled by Elhuyar.
- Two Spanish-Basque and English-Basque Machine Translation systems (MT), developed by IXA.

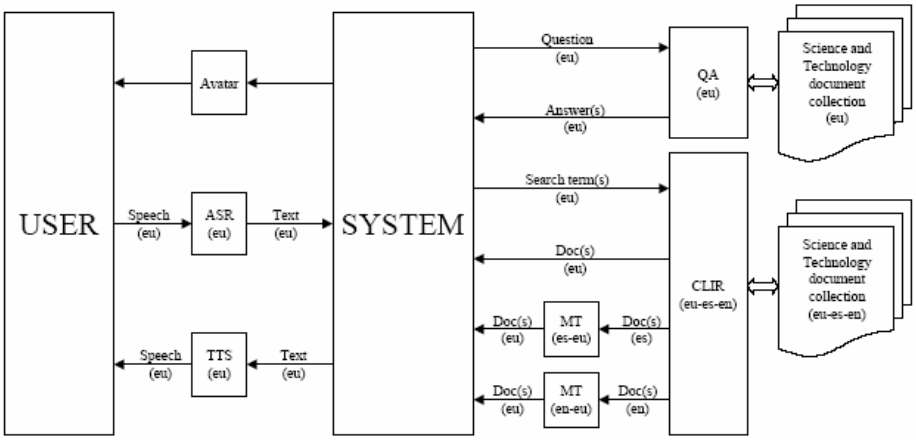

Figure 1: Diagram showing the system architecture

The AnHitz project has proved to be very effective for improving the already existing language and speech resources for Basque and for creating new ones. The system that is now being developed to integrate tools and resources from different areas (an expert in Science and Technology with a human natural language interface) shows that collaboration between agents working in different areas is crucial to really exploit the potential of language technologies and build applications for the end user.

Asistentes Virtuales Semánticos

Sonia Sánchez-Cuadrado, Mónica Marrero, Jorge Morato, Jose Miguel Fuentes[1]

Universidad Carlos III de Madrid, Departamento de Informática,
Avda. Universidad, 30, Leganés
28911 Madrid, España
{ssanchec, monica.marrero, jorge@ie.inf.uc3m.es;
[1] josemiguel.fuentes@reusecompany.com}

Resumen: Distintas técnicas de procesamiento del lenguaje permiten a sistemas automáticos simular una conversación con significado. Los agentes conversacionales o *chatbot* son diseñados para satisfacer las necesidades de información de los usuarios. A diferencia de un motor de recuperación de información tradicional, los asistentes virtuales responden con una única respuesta. Se ha desarrollado un asistente virtual basado en semántica y procesamiento del lenguaje natural para proporcionar atención de un sitio Web. Sus principales características son el uso del contexto, anticipación de consultas frecuentes y solución a consultas fuera del dominio de interrogación.

Introducción

El término asistente virtual se refiere a un sistema automático capaz de atender las necesidades de un usuario, e interactuar con él. Siguiendo esta acepción, nos encontramos con una amplia variedad de sistemas en función de la tarea que desempeñen, el grado de interacción con el usuario y su flexibilidad, e incluso la apariencia que adoptan.

Los agentes conversacionales o *chatterbots* [1] tratan de simular una conversación inteligente con uno o más personas. La conversación con un agente de este tipo se realiza mediante el reconocimiento automático de frases y/o palabras que los usuarios utilizan. Diferentes técnicas de análisis y procesamiento del lenguaje permiten a los sistemas seleccionar o elaborar las respuestas adecuadas para aparentar una conversación con significado. Por ejemplo, ante una frase como *"I am feeling very worried lately"* el agente conversacional puede estar preparado para reconocer la frase "I am" y responder remplazando *"**Why are you** feeling very worrried lately*". Esta técnica es también conocida por "el efecto de ELIZA" [2][3], por ser éste el nombre del primer sistema capaz de interaccionar con el usuario, mediante técnicas de reconocimiento de palabras clave y uso de patrones [4]. Actualmente otras técnicas utilizadas son el procesamiento de lenguaje natural (PLN), usado en los asistentes virtuales ELLA y Jabberwacky, entre otros, y el uso de técnicas de inteligencia artificial, como es el caso de A.L.I.C.E. [5].

Habitualmente, los agentes conversacionales se diseñan para tratar de satisfacer las necesidades de información de un usuario. En este sentido, su objetivo es similar al de

un sistema de pregunta-respuesta que, a diferencia de un motor de recuperación de información tradicional, tratará de responder con una única respuesta, que además suele ser un dato o texto concreto. De este modo, no obligará al usuario a navegar entre una interminable lista de resultados, en busca de la respuesta a su consulta. Habitualmente estos sistemas se centran en áreas temáticas concretas, obteniendo de este modo resultados más precisos (ej. asistente virtual de IKEA –Anna-).

Otro rasgo diferenciador de los asistentes virtuales es la interfaz, a la que cada vez se le da más importancia dada su influencia en la interacción humana. Muchos asistentes virtuales, sobre todo los de tipo comercial, muestran un personaje o avatar capaz de mostrar expresiones adecuadas a la pregunta introducida por el usuario o a la respuesta dada por el propio sistema. El objetivo es dotar de personalidad al asistente virtual, de forma que resulte más natural la interacción entre hombre y máquina. Ejemplos de este tipo de asistentes son Ella [6], Joan de Icogno [7], Vi-Clone [8], etc.

El uso de agentes conversacionales capaces de ofrecer información o de guiar a los usuarios a través de procedimientos está cada vez más extendido. Actualmente encontramos este tipo de sistemas tanto en organismos públicos como el Ministerio de Cultura, La Generalitat Catalana o La Junta de Andalucía, como en empresas privadas: Repsol, Telefónica o IKEA. También las compañías financieras los están implantando. Este es el caso de La Caixa, Bankinter, Banco Sabadell, BBVA y Caja Madrid, por citar algunos.

Asistente Virtual Semántico

El asistente virtual desarrollado, Asistente Virtual Semántico (AVS), es un agente conversacional representado por un avatar configurable, que tiene por objetivo la atención de un sitio web. Puede adaptarse a diferentes áreas temáticas, y existen tres tipos de consultas a las que puede responder:

- Cultura general: se trata de todo tipo de coletillas y frases propias de la cultura general, y que pueden ser reutilizadas en asistentes realizados para diferentes áreas temáticas o de negocio.
- Preguntas temáticas de respuesta estructurada: su respuesta viene marcada como la ejecución de una consulta a alguna base de datos. Por ejemplo, en el caso de una agencia de viajes, el sistema sabrá en todo momento contestar sobre el precio del viaje más barato a Cancún, la hora de salida de un vuelo, etc.
- Preguntas temáticas de respuesta textual: dado un conjunto de preguntas frecuentes, el sistema se encarga de encontrar cuál de todas es la que más se asemeja a la consulta del usuario.

Metodología

Tradicionalmente, los motores de indexación y recuperación documental son una pieza clave en la construcción de estos asistentes virtuales. Estos motores suelen basar su trabajo en la identificación y posterior recuperación de 'palabras' del documento donde, en muchos casos, apenas se tiene esa conciencia de 'término' y se trata cada

palabra como una secuencia de caracteres. El planteamiento para diseñar un motor de recuperación para el asistente AVS (Fig. 1), desarrollado en este trabajo, pretende obviar estos efectos. Los requisitos que debe cubrir el asistente son:

- Independencia de variantes morfológicas: tratamiento tanto de errores ortográficos, como variantes de sustantivos (singulares, plurales, masculinos, femeninos) y de verbos (tiempos y personas verbales)
- Capacidad de utilizar mapas conceptuales (ontologías): disponer de mapas que permitan indicar la cercanía semántica entre una pareja de conceptos (sinonimia, jerarquía, asociación) y permita dotar de mayor flexibilidad al sistema, aumentando el abanico de posibilidades en la respuesta.
- Capacidad de analizar documentos para identificar posibles respuestas: es posible llevar a cabo una extracción semi-automática de términos y relaciones a partir de documentos textuales, de modo que sea posible identificar posibles respuestas en caso de aparecer términos y relaciones iguales o similares.

Tras analizar los asistentes comentados al inicio del documento se consideró que la solución óptima debía basarse en Procesamiento del Lenguaje Natural y reconocimiento de patrones sintácticos.

De este modo, el sistema debe satisfacer, a grandes rasgos y dada una consulta, los siguientes pasos:

1. Acotación del dominio: qué tipología de preguntas deberá responder
2. Tokenización
3. Corrección ortográfica
4. Lematización
5. Desambiguación y agrupación semántica de los verbos y sustantivos.
6. Generación de un patrón basado en relaciones

Una vez realizados estos pasos, el sistema contrasta la información contenida en la base de conocimiento del AVS y determina y/o elabora la respuesta idónea y más relevante.

Otras características necesarias para un buen rendimiento del asistente son:

- Uso del Contexto
- Anticipación de consultas frecuentes y solución a consultas fuera del dominio de interrogación.

Uso del contexto

Al igual que en las conversaciones con humanos, las conversaciones con asistentes virtuales pueden requerir la memorización de ciertos datos. Por tanto, el asistente virtual debe poder recordar características de las consultas anteriores para incluirlas a modo de contexto en la resolución de una nueva consulta. El AVS es capaz de incorporar a las preguntas de los usuarios el contexto de la pregunta inmediatamente anterior. Un ejemplo sería la siguiente conversación:

[Usuario]: ¿Disponéis de la titulación de Informática de Gestión?
[Asistente]: Sí, así es. Tienes más información en la siguiente URL….
[Usuario]: y Estadística

[Asistente]: Sí, entre nuestras asignaturas contamos con la de Estadística. Puedes …
[Usuario]: y cuál es el plan de estudios
[Asistente]: El plan de estudios de la titulación de Estadística es el siguiente….

Cuando el usuario pregunta 'y estadística' el Asistente ha sido capaz de recordar que se está preguntando sobre si se dispone información de esa titulación. Igualmente, al preguntar 'y cuál es el plan de estudios', el Asistente recuerda que estamos hablando de la titulación de Estadística y ofrece el plan de estudios de la misma.

El uso del contexto puede aportar respuestas a preguntas que de otro modo no podrían ser respondidas, sin embargo, su uso debe ser muy limitado ya que de lo contrario se introduciría ruido. Por este motivo se utiliza tan sólo cuando no es posible localizar una respuesta de otro modo y teniendo en cuenta únicamente la pregunta inmediatamente anterior.

Fig. 1 Asistente Virtual Semántico

Conclusiones

Los agentes conversacionales dirigidos a aportar información al usuario o guiarlos en procedimientos son sistemas complejos que requieren de la participación de diversas disciplinas. Quizá los sistemas más parecidos son los Sistemas de Pregunta-Respuesta (QA), donde el procesamiento del lenguaje natural, los patrones y el uso de heurísticas o incluso modelos de aprendizaje son técnicas habituales. Sino que además es necesario dotarlos de conocimiento general, temático y de contexto, de manera que sean capaces de responder a preguntas o aseveraciones propias del lenguaje común (ej. Hola), a preguntas específicas del área temática que se espera que conozcan (ej. Procedimientos bancarios en entidades financieras) e incluso a preguntas relacionadas con otras anteriores. La personalización y expresividad, donde se aúnan la inteligencia artificial y la psicología, también ayuda a lograr fluidez en las conversaciones.

Por otro lado, los beneficios de un asistente virtual son elevados. Esto se debe a la disponibilidad espacial y temporal que puede presentar de cara a los usuarios,

supliendo a comerciales o expertos, pero también a que, desde un punto de vista comercial, los asistentes virtuales pueden dar a los gestores de las compañías que los implementan, valiosísima información como por ejemplo:

- Estadísticas de uso del sistema: número de consultas por día, por horas, totales…
- Cuáles son las consultas más frecuentes, e incluso los términos más utilizados

Esta información permite valorar el éxito de una deteminada acción de marketing, o descubrir necesidades de nuestros clientes en las que aún no se había reparado, ofreciendo además una realimentación necesaria para la mejora del propio asistente.

Bibliografía

[1] Mauldin, Michael L CHATTERBOTS, TINYMUDS, and the Turing Test: Entering the Loebner Prize Competition. AAAI, 1994. pp. 16-21

[2] Hofstadter, Douglas R. "Preface 4 The Ineradicable Eliza Effect and Its Dangers,Epilogue". Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought. Basic Books, 1996. p. 157. http://books.google.com/books?id=somvbmHCaOEC&pg=PA157

[3] Fenton-Kerr, Tom, "GAIA: An Experimental Pedagogical Agent for Exploring Multimodal Interaction", Computation for Metaphors, Analogy, and Agents, Springer, 1999. p. 156

[4] Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1. 1966. pp. 36-45. DOI= http://doi.acm.org/10.1145/365153.365168

[5] Wallace, R.S. The Anatomy of A.L.I.C.E. A.L.I.C.E. Artificial Intelligence Foundation, Inc. 2004. http://www.alicebot.org/documentation/

[6] Ella Systems, 2009. http://www.ellaz.com/AIV/default.aspx

[7] Joan - Our Artificially Intelligent, speaking, videocentric Avatar. 2009 http://www.icogno.com/

[8] Web de Vi-Clone http://www.vi-clone.com/es/

Thuban: Acompañamiento Virtual mediante Dispositivos Móviles e Interacción Natural

Dolores Cuadra[1], Javier Calle[1], David del Valle[1], Jessica Rivero[1]

[1] Universidad Carlos III de Madrid, Departamento de Informática
28911 Leganés, España
{dcuadra, fcalle, dvalle, jrivero}@inf.uc3m.es

Abstract. Este proyecto persigue la realización de un acompañante virtual capaz de interaccionar con el usuario mientras este se mueve en un entorno, proporcionándole acceso a un conjunto de servicios preestablecidos. Entre estos servicios figuran aquellos que tienen en cuenta la posición y trayectoria del usuario (p.e., avisos), los orientados a dirigir esos parámetros (p.e. establecimiento de rutas y guiado hacia puntos de interés fijos o móviles) o explicarlos (descripción de situación y/o trayectoria), y en general cualquier servicio disponible en el sistema (p.e., búsquedas en catálogo). La interacción del sistema con el usuario deberá estar orientada a la imitación de la interacción humana (interacción natural) de modo que cualquier persona no entrenada tecnológicamente pueda disfrutar de la facilidad desarrollada. Esta interacción se desarrolla gracias a la intervención de diversos modelos de conocimiento, entre los que cabe destacar el de Diálogo (con procesamiento intencional de acción combinada) y el de Situación. Este último se apoya en la tecnología de Bases de Datos Espacio Temporales, capaz de manejar el posicionamiento de elementos móviles y de gestionar eventos para situaciones espacio-temporales. La integración de estas tecnologías es la base de la plataforma de interacción que posibilita una generación de Acompañantes Virtuales en entornos culturales y de interés social, o meramente comerciales (realidad aumentada).

Keywords: Interacción Natural, Bases de Datos Espacio-Temporales, Modelos de Diálogo, Modelos de Situación, Sistemas Multiagente, Realidad Aumentada

1 Introducción

Durante los últimos años, en el área de la Interacción Persona-Ordenador (IPO) ha sido creciente el interés en la interacción en la que participan usuarios no entrenados tecnológicamente. La única habilidad interactiva del usuario es la que le permite interaccionar con otros individuos de su misma especie, es decir, otros humanos. En definitiva, se persigue imitar a los seres humanos en su manera de interactuar [1], y la disciplina que lo pretende recibe el apelativo de Interacción Natural (IN).

A finales de la década de los noventa, algunos investigadores coincidían al asegurar que era el momento de embarcarse a descubrir cómo procesar aquello que "se profiere de modo natural" [13]. La concienzuda recopilación de Clark [6] dio

solidez y completitud a las bases teóricas que habrían de servir de guía a los sistemas que pretendan 'utilizar el lenguaje humano'. Algunas de las características que se esperan en sistemas de este tipo ya habían sido anticipadas por algunos sistemas de interacción, los que Cohen [7] define como 'sistemas de interacción enfocados como una actividad combinada'. Cabe destacar en este sentido los del equipo de Allen en Rochester [9], que además de destacar en otros aspectos, introducen e implementan conceptos orientados a una interacción natural.

2 Interacción Natural

Un problema de base para la consecución de este paradigma de interacción es que el origen de cada expresión puede ser diverso: puramente interactivo, emocional, operativo (referente a la tarea), etc. Si no pudiera considerarse alguno de estos factores (por su elevada complejidad y coste) se prescindiría de sus efectos sobre la interacción, mermando su naturalidad (produciendo comportamiento mecánico), pero este planteamiento parcial si es adecuado en investigación para avanzar en algunos de los factores. Por otro lado, ha tomado un auge importante en la IPO el enfoque del diseño basado en modelos, que sienta las bases de sistemas interactivos que contemplen todos los tipos de conocimiento que influyen en una expresión.

Debido a la diversidad en la naturaleza de ese conocimiento, que implica distintas maneras de estructurarlo y manejarlo, se hace preciso clasificarlo y distribuirlo entre diversos modelos de conocimiento especializados (estructuras de conocimiento más mecanismos de razonamiento sobre el conocimiento específico que conduzcan a las decisiones adecuadas en cada momento de la interacción). De este modo, pueden surgir distintas Aproximaciones (o arquitecturas) Cognitivas [10]. Cada uno de sus modelos seguirá posteriormente un proceso propio de ingeniería del conocimiento, adaptado al tipo de conocimiento y que aplique los avances propuestos en su área.

Existen diversas propuestas a este respecto, entre las que cabe destacar las del equipo de Allen en Rochester [2], o la de Chai en IBM [5]. Otra propuesta más reciente encabezada por Wahlster [14] abre camino a los sistemas IN basados en modelos de conocimiento específico que son desarrollados y evolucionan de modo independiente. Estos sistemas se vienen implementando sobre plataformas multi-agente que posibilitan el funcionamiento autónomo y concurrente de cada modelo.

En el grupo en el que se va a desarrollar esta propuesta (grupo LABDA, de la Universidad Carlos III de Madrid), se sigue la aproximación presentada en [3], basándose en la experiencia obtenida en numerosos proyectos de investigación de ámbito europeo y nacional. En esta propuesta tienen un protagonismo destacado los Modelos de Diálogo, Situación, Usuario, y Ontología, así como los componentes de interfaz y el modelo de Presentación.

El Modelo de Diálogo estructura y maneja el conocimiento referente a las secuencias de acciones comunicativas que forman la interacción, así como su organización y ordenación en el tiempo y las causas que las originan. Deberá procurar una interacción flexible, coherente y natural. Las aproximaciones clásicas a este tipo de modelado se centran en validar la estructura del diálogo, mediante Gramáticas o Juegos de Diálogo [12], y en planificar los caminos que llevarán a estados

satisfactorios del diálogo [15], que sí garantizan la coherencia pero se caracterizan por su rigidez y mecanicidad (producen diálogos mecánicos, no naturales). Para usar el lenguaje de un modo natural, un paso ineludible según Clark es tener en cuenta que la interacción es una actividad combinada entre dos (o más) participantes. El sostenimiento de una zona común y el compromiso de ambos en cada meta como prerrequisito para su desarrollo son aspectos cardinales en este tipo de actividad.

El Modelo de Hilos [3] aplica algunas de estas teorías con el fin de involucrar al sistema en la búsqueda del 'mantenimiento del compromiso y de la sintonía' entre usuario y sistema, proponiendo un procesamiento intencional que permite introducir técnicas de reparación y refuerzo del compromiso de las metas compartidas por los interlocutores, cuando este se debilita, con el fin de mantener la sintonía entre ambos. Su enfoque separa los aspectos intencional (dinámico), estructural, y contextual (estático), subordinando estos dos últimos a la estructura del primero. Para la organización intencional, diferencia tres espacios: el del usuario, el del sistema, y una zona común o compartida por ambos. Se caracteriza por permitir la introducción en el diálogo de iniciativas del sistema (originadas en el modelo de diálogo o en cualquier otro modelo de la arquitectura), y por contemplar las interrupciones (robos de turno) que ocurren en la IN (diálogo de turnos solapados).

El Modelo de Situación agrupa su conocimiento en cinco categorías o aspectos según Gee [11]: semiótico (signos que se usan en la interacción), operativo (tarea/s que subyace/n a la interacción), material (características espacio temporales), político (roles de los participantes), y socio-cultural. Los servicios que proporciona se orientan en tres direcciones: (a) identificar la situación (teniendo en cuenta todos o algunos de sus aspectos) de modo que otros modelos pueden filtrar su propio conocimiento (para aplicar solo aquel que sea relevante en la situación dada); (b) programar y disparar eventos por la situación y sus consecuencias, de modo que se realicen cuando se alcanzan determinadas circunstancias; y (c) proporcionar conocimiento relativo a la situación (circunstancias pasadas, predicciones sobre circunstancias futuras, o secuencias de acciones para alcanzar una circunstancia partiendo de otra, por poner unos ejemplos). Atendiendo a las características de este modelo, se hace preciso contar con una tecnología que soporte sus necesidades: las BD espacio-temporales.

3 Plataformas multi-agente para la Interacción Natural

Tal y como se ha descrito anteriormente, la tendencia en el desarrollo de este tipo de sistemas se encamina a obtener modelos de conocimiento específicos que operen simultáneamente de modo autónomo. Para llevar a cabo esta aproximación, la solución que mejor se ajusta vendrá de la mano de un sistema multi-agente [8][4].

Algunos de los beneficios que este tipo de sistemas aportan a la IN son: actualización simultánea del estado de cada modelo de conocimiento (en su caso) a medida que se desarrolla la interacción; procesamiento continuo (versus procesamiento clásico por turnos); producción de distintas soluciones basadas en la cooperación de distintos tipos de conocimiento y selección de la más adecuada en tiempo real; evolución independiente de cada modelo, facilitando la reutilización

individual de los componentes y la escalabilidad del sistema; fácil incorporación de agentes externos y aplicaciones, que permite integrar la IN casi en cualquier sistema.

Referencias

1. Bernsen, N.O. What is Natural Interactivity? In Procs. of WS: From Spoken Dialogue to Full Natural Interactive Dialogue. Theory, Empirical Analysis and Eval., pp. 34-37. Ed. L. Dybkjær. 2nd Int. Conf. on Language Resources and Evaluation (LREC'2000).
2. Blaylock,N., Allen,J., Ferguson,G. Managing Communicative Intentions with Collaborative Problem Solving. Jan van Kuppevelt and Ronnie W. Smith (eds.), Current and New Directions in Discourse and Dialogue, pp. 63-84. © 2003 Kluwer Academic Publishers.
3. Calle, J., García-Serrano, A., Martínez, P. Intentional Processing as a Key for Rational Behaviour through Natural Interaction. To appear in Interacting With Computers, © 2006 Elsevier Ltd.
4. Calle, J., Martínez, P., Valle, D., Cuadra, D. Towards the Achievement of Natural Interaction. In: Engineering the User Interface: from Research to Practice, Springer 2008.
5. Chai, J. Pan, S., Zhou, M. MIND: A Context-based Multimodal Interpretation Framework in Conversational Systems. In Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, Eds. O. Bernsen , L. Dybkjaer and J. van Kuppevelt; Kluwer Academic Publishers, 2003.
6. Clark, H.H. Using Language. © 1996, Cambridge University Press.
7. Cohen, P.R., 1997. Dialogue Modeling. In Survey of the state of the art in Human Language Technology; chap.6, pp. 204-209. Cambridge University Press, 1998.
8. Cuadra D., Rivero J., Valle D., Calle F.J. Enhancing Natural Interaction with Circumstantial Knowledge. Int. Trans. on Systems Science and Applications (ISSN 1751-1461), Vol 4, No. 1 Special Issue on Agent based System Challenges for Ubiquitous and Pervasive Computing, 2008.
9. Ferguson, G.F., Allen, J.F. TRIPS: An Integrated Intelligent Problem-Solving Assistant. In Procs. of the 15th National Conference on Artificial Intelligence, 1998; pp 567-572.
10. Garcia-Serrano, A., Calle-Gómez, J. A cognitive Architecture for the design of an Interaction Agent. Cooperative Information Agents VI. Editors M. Klusch, S.Ossowski & O. Shehory. Lecture Notes in Artificial Intelligence, pp. pp 82-89; Springer 2002.
11. Gee, J.P. Introduction to Discourse Analysis. Routledge, 1999.
12. Levin, J. A. and Moore, J. A. 1977. Dialogue games: Metacommunication strategies for natural language interaction. Cognitive Science, 1(4):395–420.
13. Oviatt,S.L., Cohen, P.R. Multimodal Interfaces That Process What Comes Naturally. Communications of the ACM 43(3): 45-53 (2000).
14. Wahlster W., et al. SmartKom: foundations of multimodal dialogue systems. Springer (2006).
15. Young, R. M., Moore, J. D., Pollack, M. E. Towards a Principled Representation of Discourse Plans. In procs. of the Sixteenth Conference of the Cognitive Science Society , Atlanta, GA, 1994.

A Model for Representing and Accessing Web Services through a Dialogue System

Meritxell Gonzàlez[1] and Marta Gatius[1]

TALP Research Center, Campus Nort UPC
Jordi Girona 1-3, 08034 Barcelona
gonzalez,gatius@lsi.upc.edu,
http://www.lsi.upc.es/~nlp/disi/

Abstract. Dialogue Systems can be used for guiding the user accessing the web. We present a model for representing the information related to a web service, and the algorithms for processing the data. In our work, models and algorithms are domain and application independent, and the application specific resources are provided accordingly. The data models contain the information related to current interaction with the system. They are classified along with a typology of tasks. The task definitions states the specific application information needed for executing the web service. Algorithms matches the information from the data model against the task definition, generating the execution of the web service and the information that the dialogue manager needs to make inquires about the interaction.

1 Introduction

The use of Natural Language can also increase web services usability and accessibility. Dialogue Systems (DSs) can be used for guiding the user accessing web applications. However, DSs are usually developed for a specific type of application, and the adaptation of existing DSs to new applications becomes a hard task. In this area, we can currently found some research investments on empowering data models and algorithms for developing DSs that are reusable for several types of application [3],[1].

Our effort in this direction is focused on developing data models and resources that can be easily adapted to any application being integrated in our DS [2]. Figure 1 represents the architecture of our system.

In this work we describe a model for representing the type of tasks available through Web Services, as well as a data structure for the application's resources. We finally give an overview of the algorithms that process that information. These algorithms generate the data used by the Dialogue Manager for planing its interventions.

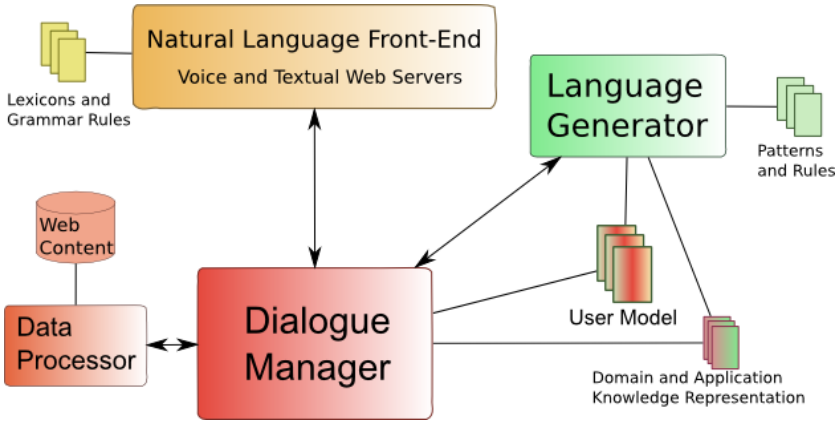

Fig. 1. Dialogue System Architecture

2 Tasks Tipology and Data Model

We consider that a web service is related to one application, and that each application can perform one or more operations. Each operation is later divided into tasks, that are further related to dialogue actions. Tasks can be shared among different operations of a web service, but they cannot be shared among different web services. We consider that tasks in web services can mostly fall in any of these three types of applications: Transactions, Queries and Searches.

The Task Data Model (TDM) is the information related to any task that is being executed in the DS. The required information and its structure depends on which type of task it belongs to.

Search task is the most commonly used operation in the web: to search and get a list of things given some restrictions or requirements. We call these restrictions the *queryConstraints* of the Search, and the *requestedData* is the specification of the information the user is looking for. For example, listing all the cinemas (*requestedData*) where a specific film (*queryConstraints*) is played. We classify the results of a Search task in two main groups depending if there are results or changes in the constraints are required. In case that there are results, we distinguish between a list of objects and a single object. In the other hand, if changes in the constraints are needed, we distinguish between stepping up constraints and relaxing them.

Query task stands for searching a concrete object given a *queryConstraints*, as in the Search task. However, the result of a Query task is not a list of objects, but a list of features or characteristics related to an unique and identifiable object in the web service. For example, all the information related to an specific cinema: the address, the services, the timetable, etc.

Transaction Tasks are those that given some parameters' values, they perform the confirmation of the values and then executes the application transaction

related to the Web Service. These Task could eventually return any type of information. For example, booking an hotel returns the bono-document with the reservation.

3 Tasks Definition

In order to process the information contained in the TDM, the system needs to know the Tasks Definitions (TDs) of the web service. TDs are developed when integrating a new application in the DS. They consists of (i) the operations and tasks that the web service performs, (ii) the sets of input parameters for each task, (iii) the information that each task returns, and (iv) further requirements and conditions among and related to all the involved data.

The process of executing a specific task consists of instantiating a specific TD, chosen accordingly with the information contained in the TDM.

Search definition states the allowed sets of *queryConstraints* and *requestedData*, as well as possible default values and their conditions of use (for example, *queryConstraints* shouldn't include values stated in the *requestedData*). The TD can also include directions for generating summaries and relaxation of the data constraints, for being used after the execution.

Query definition states the allowed set of *queryConstraints* attributes. Any of these available set of *queryConstraints* must be enough for identifying an unique entity in the web service. The TD states as well conditions and constraints among the attributes as well as constraints overall the execution. It also states the type of the object that the web service will eventually return.

The Transaction is more intuitive and similar to a common application procedure. Its TD states a list of input parameters (some of them are mandatory and others are optional). As in the other TDs, it can define constraints about the filled data, as well as the execution conditions.

4 Algorithms

The algorithms used for processing the information related to the current task in execution and its definition are called Recipes. The information from the TD is matched against the information contained in the TDM.

The Search recipe is the algorithm for processing the information related to a Search task. First, all *queryConstraints* conditions are evaluated. If none of them is suitable, then relaxation rules if available are used. If at least one suitable *queryConstraints* is found, then the *requestedData* is evaluated. If it is not suitable, then it is updated accordingly the *queryConstraints* information and the default values defined in the task. Once the *requestedData* is also suitable,

then the Search can be executed. The result of the execution is also classified following the four types of results considered in the system.

The Query recipe is the algorithm that processes Query tasks. It also evaluates firstly the conditions in the *queryConstraints*. Then, as soon as one suitable *queryConstraints* is found, the Query is executed. Evaluating successfully the whole data is not needed. The result of the execution must be an unique object, or nothing. In case the system finds more than one object, it shows there is something wrong in the resources or in the TD.

The Transaction recipe algorithm consists of evaluating the mandatory parameters, and then evaluate optional ones, removing the non suitable. Since Transactions updates the systems databases, the parameters used in the Transaction must be confirmed by the user that is using the DS. It is a task of the DS to obtain the data evidences stating that the Transaction parameters have been confirmed. Finally, the Transaction execution returns the transaction status, and, just in case, an object from the web service.

References

1. N. Blaylock. *Towards Tractable Agent-based Dialogue*. PhD thesis, University of Rochester, Rochester, New York, 2005.
2. M. Gatius, M. González, and E. Comelles. An information state-based dialogue manager for making voice web smarter. In *Proceedings of the 16th International Worl Wide Web Conference*, pages –, 2007.
3. J. Schehl, S. Ericsson, C. Gerstenberger, and P. Manchón. Plan library for multimodal turn planning. European Project TALK (IST-507802) Deliverable D3.2, 2007.

Enriching ontologies with multilingual information

Guadalupe Aguado de Cea[1], Asunción Gómez-Pérez[1], Elena Montiel-Ponsoda[1]

[1] Ontology Engineering Group, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid
{lupe, asun, emontiel}@fi.upm.es

Abstract. Multilinguality in ontologies has become an impending need for institutions worldwide that have to deal with data and linguistic resources in different natural languages. Since most ontologies are developed in one language, obtaining multilingual ontologies implies to *localize* or adapt them to a concrete language and culture community. As the adaptation of the ontology conceptualization demands considerable efforts, we propose to modify the ontology terminological layer by associating an external repository of linguistic data to the ontology. With this aim we provide a model called *Linguistic Information Repository* (LIR) that associated to the ontology meta-model allows terminological layer localization.

Keywords: multilingual ontologies, Ontology Localization, Linguistic Information Repository (LIR)

1 Introduction

Multilinguality in ontologies is nowadays demanded by institutions worldwide with a huge number of resources in different languages. One of these institutions is the FAO[1]. Within the NeOn project[2], the FAO is currently leading a case study on fishery stocks in order to improve the interoperability of its information systems. The FAO, as an international organization with five official languages -English, French, Spanish, Arabic and Chinese- deals with heterogeneous and multilingual linguistic resources with different granularity levels. This scenario is an illustrative example of the need for semantically organizing great amounts of multilingual data. When providing ontologies with multilingual data, one of the activities identified in the NeOn ontology network development process is the Ontology Localization Activity, that consists in *adapting an ontology to a concrete language and culture community*, as defined in [1]. In particular, our aim is to obtain multilingual ontologies by localizing its terminological layer (terms or labels that name ontology elements), rather than modifying its conceptualization. Thus, we propose to link ontologies with a linguistic model, called *Linguistic Information Repository* (LIR), whose main feature is that it provides (1) a complete and complementary amount of linguistic data

[1] http://www.fao.org/
[2] http://www.neon-project.org/

that allows localization of ontology elements to a specific linguistic and cultural universe, and, (2) a unified access to aggregated multilingual data.

2 Related Work

Regarding the activity of Ontology Localization, we have identified three ways of modelling multilinguality in ontologies: 1) inclusion of multilingual labels in the ontology by means of the `rdfs:label` and `rdfs:comment` properties (most widespread modality), 2) mapping of several conceptualizations in different natural languages through an interlingual set of common concepts (as in EWN[3]), and 3) association of an external linguistic model to the ontology (as in LingInfo [2]). The first modality option restricts the amount and type of linguistic information that can be associated to the ontology. The second option requires a huge effort at two stages: first, when a new language has to be integrated in the multilingual system, since a new conceptualization has to be developed, and second, by the establishment of alignments among conceptualizations or between the new conceptualization and the interlingua. In this way, our hypothesis is that the best solution lies on the third option, in which the type and quantity of linguistic information is not restricted, and the linguistic elements that compose the model can be related among them. Regarding this latter option, we argue that existing models have not been intended to cover localization needs, and do not include enough information in this sense, but rather focus on other linguistic information such as the morphosyntactic aspects of ontology labels.

3 Proposed Approach

With the aim of providing available ontologies in one natural language with multilingual information thus allowing their localization, we have designed the LIR [3,4,5] within the NeOn project. The LIR is an external linguistic model based on existing linguistic (LMF[4]) and terminological (TMF[5]) representation schemas. The LIR permits the association of a set of linguistic data to any element in the ontology. The main classes or data categories that compose the LIR are: *LexicalEntry*, *Lexicalization*, *Sense*, *Definition*, *Language*, *Source*, *Note*, and *Usage Context* (as can be seen in Figure 1). Thanks to the relations that can be established among the LIR classes, the LIR mainly accounts for: well-defined relations within lexicalizations in one language and across languages, and conceptualization mismatches among different cultures and languages. The main benefits of this approach against the modeling options presented in section 2 are the following: a) the association of an unrestricted quantity of linguistic information to ontology elements; b) the establishment of relations among the linguistic elements, as well as the performance

[3] http://www.illc.uva.nl/EuroWordNet/
[4] Lexical Markup Framework ISO/CD 24613
[5] Terminological Markup Framework ISO 16642

of complex operations (reasoning) with them; c) the access and manipulation of the linguistic data (terminological layer) without interfering with the conceptualization, with the resulting benefits for non-ontology engineers; and d) the reuse of the contained linguistic information for other applications.

Up to now, the LIR has been implemented as an ontology in OWL [4], and is currently supported by the LabelTranslator NeOn plug-in [6] for an automatic localization of ontologies. A first set of tests has been conducted within NeOn to asses the suitability of the LIR model for the linguistic needs of the FAO. The LIR has proved to satisfy the FAO needs for i) establishing relations among lexicalizations within and across languages, ii) specifying variants for dialects or local languages, and iii) explicitly expressing translation specificities.

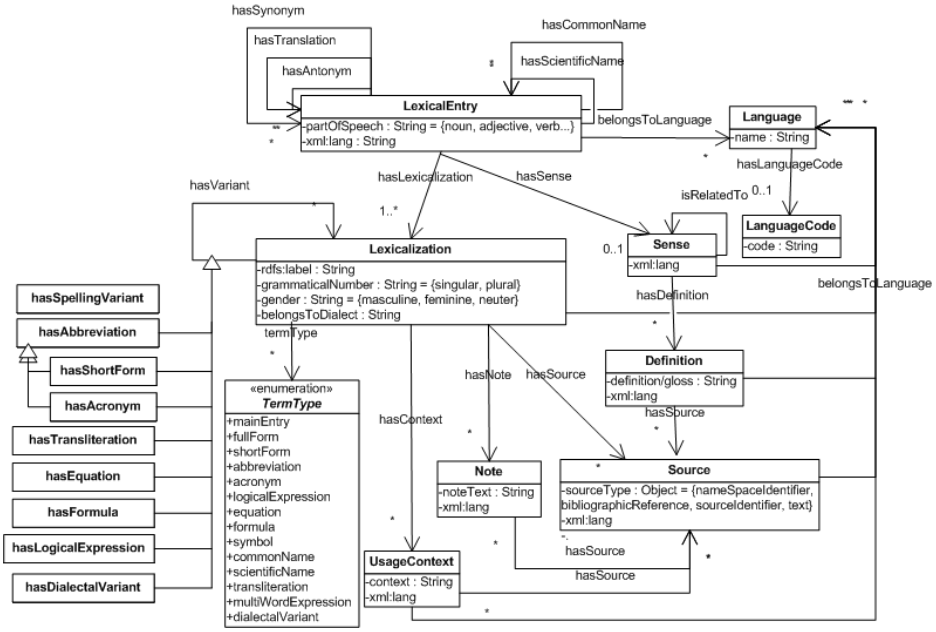

Figure 1. The LIR Model

Main References

1. Suárez-Figueroa, M.C. and Gómez-Pérez, A. *First Attempt towards a Standard Glossary of Ontology Engineering Terminology*. 8th International Conference on Terminology and Knowledge Engineering (TKE2008), Copenhagen (2008)
2. Buitelaar, P., M. Sintek, M. Kiesel. 2006. *A Multi-lingual/Multimedia Lexicon Model for Ontologies*. In Proc. ESWC'06, Budva, Montenegro (2006)
3. Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., and Peters, W. Modelling multilinguality in ontologies. In *Coling 2008: Companion volume - Posters and Demonstrations*, Manchester, UK (2008)

4. Montiel-Ponsoda, E. and Peters, W. (coordinators): Multilingual and Localization Support for Ontologies. NeOn Project Deliverable 2.4.2 (2008)

5. Peters, W., Montiel-Ponsoda, E., Aguado de Cea, G. *Localizing Ontologies in OWL*. In Proc. OntoLex, Busan, South Corea (2007)
6. Espinoza, M., Gómez-Pérez, A., Mena, E.: Enriching an Ontology with Multilingual Information. In Proc. of ESWC'08, Tenerife (Spain), LNCS Springer, ISBN 978-3-540-68233-2, ISSN-0302-9743, pp. 333--347 (2008)

Automatic Localization of Ontologies with LabelTranslator

Mauricio Espinoza[1], Asunción Gómez-Pérez[1], Eduardo Mena[2]

[1] Ontology Engineering Group, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid
asun@fi.upm.es, mespinoza@delicias.dia.fi.upm.es
[2] IIS Department, Universidad de Zaragoza, María de Luna 1, 50018 Zaragoza
emena@unizar.es

Abstract. Organizations working in a multilingual environment demand multilingual ontologies, but these are nearly inexistent in the Web. To solve this problem we propose LabelTranslator, a system that takes as input an ontology whose labels are described in a source natural language and obtains the most probable translation of each label in a target natural language. Our main contribution is the automatization of this process which reduces the human efforts implied in the manual localization of ontologies. For this aim, LabelTranslator relies on available translation services and multilingual resources, and sorts out translation senses according to similarity with the lexical and semantic context of each ontology label.

Keywords: Ontology Localization, multilingual ontologies, LabelTranslator

1 Introduction

Currently, more and more organizations working in multilingual environments demand ontologies supporting different natural languages. In the framework of the NeOn project[1], all case studies have expressed the need for multilingual ontologies. One case study is led by the Food and Agriculture Organization of the United Nations (FAO), an international organization that manages information in more than fifteen languages. The second use case is concerned with the pharmaceutical industry in Spain, and requires ontologies in the different languages spoken in the country. These are just two illustrative scenarios of the nowadays impending need for multilingual ontologies. With the aim of solving this problem, we propose LabelTranslator, a NeOn plugin that automatically localizes ontologies in English, Spanish and German. We understand Ontology Localization as the activity that consists in adapting an ontology to a concrete language and culture community, as defined in [1]. In this contribution we describe the main functionalities of the current prototype of our system.

[1] http://www.neon-project.org

2 LabelTranslator fuctional overview

LabelTranslator [2,3] has been designed with the aim of automating ontology localization, and has been implemented in the ontology editor NeOn Toolkit as a plugin. In its current version, it can localize ontologies in English, German and Spanish. In the following, we briefly describe the main tasks followed by the system in performing the localization activity.

Once an ontology has been created or imported in the NeOn ontology editor, LabelTranslator allows users and domain experts to manually sort out the ontology elements that should undergo localization. By default the system selects the whole ontology. For each ontology element, LabelTranslator retrieves its *local context* (set of hypernym, hyponym, attributes, and sibling label names associated with the term under consideration), which is interpreted by the system using a *structure-level* approach.

In order to obtain the most appropriate translation for each ontology element in the target language, LabelTranslator accesses multilingual linguistic resources (EuroWordNet[2], Wiktionary[3], or IATE[4]) and translation web services (GoogleTranslate[5], BabelFish[6], etc.) available on the Web. From these resources, the system obtains equivalent translations for all selected labels. Then, it retrieves a list of semantic senses for each translated label, querying remote lexical resources as EuroWordnet or third-party knowledge pools such as Watson[7], which indexes many ontologies available on the Web. Finally, the senses of each context label are as well discovered following the strategy just explained. At this point, it should be noted that LabelTranslator includes a compositional method to translate compound labels, which first searches for translation candidates of each token of the compound label, and then builds the translations for the candidates using lexical templates. For a detailed explanation see [2].

Then, the system uses a disambiguation method to sort out the translations according to their context. LabelTranslator carries out this task in relation to the senses of each translated label and the senses of the context labels. At this stage, domain and linguist experts may decide to choose the most appropriate translation from the ones in the ranking. In default of this, the system will consider the one in the highest position.

The ontology is updated with the resulting linguistic data, which is stored in the LIR model [4], a separate module adopted by the LabelTranslator NeOn plugin for organizing and relating linguistic information within the same language and across languages for domain ontologies.

[2] http://www.illc.uva.nl/EuroWordNet/
[3] http://en.wiktionary.org/wiki/
[4] http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load
[5] http://www.google.com/translate_t
[6] http://babel¯sh.altavista.com/
[7] http://watson.kmi.open.ac.uk/WatsonWUI/

3 Related work

Our work enhances the work presented in [5], where a system for supporting the multilingual extension of ontologies expressed in just one natural language was proposed. This tool was used to support “the supervised translation of ontology labels". Therefore, the tool offers a semi-automatic strategy. In our approach we have implemented an automatic method to reduce human intervention while enriching an ontology with linguistic information.

In [6] the authors propose a framework for adding linguistic expressivity to conceptual knowledge, as represented in ontologies. They use two lexical resources for linguistic or multilingual enrichment: WordNet, and DICT dictionaries. In this work, the process to translate compound ontology labels is not described.

In [7] a method to give support to multilingual ontology engineering is developed. In this work some software tools have been used for supporting the process of term extraction and translation. In particular, the translation process requires sentence aligned parallel text, tokenized, tagged and lemmatized. In our opinion, obtaining a corpus aligned is not a simple task. Unlike this work, we rely on some multilingual translation services and extend them by using lexical templates.

References

1. Suárez-Figueroa, M.C. and Gómez-Pérez, A. First Attempt towards a Standard Glossary of Ontology Engineering Terminology. In *Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE2008)*, Copenhagen (2008)
2. Espinoza, M., Gómez-Pérez, A., Mena, E. Enriching an Ontology with Multilingual Information. In *Proceedings of 5th European Semantic Web Conference (ESWC'08)*, Tenerife (Spain), LNCS Springer, ISBN 978-3-540-68233-2, ISSN-0302-9743, pp. 333--347 (2008)
3. Espinoza, M., Gómez-Pérez, A. and Mena, E. LabelTranslator - A Tool to Automatically Localize an Ontology. In *Proceedings of 5th European Semantic Web Conference (ESWC'08)*, Tenerife (Spain), Springer Verlag LNCS, ISBN 978-3-540-68233-2, ISSN-0302-9743, pp. 792-796, demo paper (2008)
4. Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., and Peters, W. Modelling multilinguality in ontologies. In *Coling 2008: Companion volume - Posters and Demonstrations*, Manchester, UK (2008)
5. Declerck, T., Gómez-Pérez, A., Vela, O., Gantner, Z. and Manzano-Macho, D. Multilingual lexical semantic resources for ontology translation. In *Proceedings of LREC 2006* (2006)
6. Pazienza M.T., and Stellato, A. Exploiting linguistic resources for building linguistically motivated ontologies in the semantic web. In *Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006), held jointly with LREC2006, May 24-26, 2006, Genoa, (Italy)*, 2006.
7. Kerremans, K., and Temmermann, R. Towards multilingual, termontological support in ontology engineering. In *Proceedings Workshop on Terminology, Ontology and Knowledge representation*, pp. 22-23, Lyon, France (2004)

Etiquetado Semántico De Notas Clínicas Sobre SNOMED

Elena Castro, Leonardo Castaño

[1] Computer Science Department of the University Carlos III of Madrid, Spain
{ecastro, lcastano}@inf.uc3m.es

Abstract. El procesamiento de información médica relativa a historiales y notas clínicas es una tarea ardua debido a la elaboración manual de este tipo de información y a la diversidad de terminología que contienen. En esta contribución se presenta una herramienta de reconocimiento de conceptos médicos en español utilizando el meta-tesauro SNOMED.

Keywords: etiquetado semántico, meta-tesauro, SNOMED.

1 Introducción

El procesamiento automático de textos médicos es una de las áreas que están cobrando especial interés en los últimos años, debido a varios factores: en primer lugar la gran cantidad de material científico producido, unido a la necesidad de herramientas automáticas de consulta y gestión, y por último, a la dificultad de procesamiento de información elaborada por especialistas en diversas áreas de la medicina en sus historiales y notas clínicas. Esta información consiste en la mayoría de las ocasiones en registros que contienen datos no estructurados, usualmente elaborados de forma manual (lo que puede inducir a omisión de determinadas reglas de ortografía como acentos. ….) y sin seguir una única convención en cuanto a transcripción de conceptos, escritura y colocación de abreviaturas, etc., además del grave problema para la protección de datos que supone el que se incluya información de carácter personal (nombres de personas, entidades, …) que podría revelar datos confidenciales de pacientes y/o especialistas.

Centrándose en el procesamiento de notas clínicas, en lengua inglesa se han elaborado diferentes recursos como MeSH, UMLS, etc. [1], sin embargo, en otros idiomas como el español se echa en falta este tipo de recursos. En este marco y dentro del proyecto ISSE (FIT-350300-2007-75), este trabajo presenta una herramienta que basada en el meta-tesauro SNOMED para terminología médica, y más concretamente en su subconjunto de descripciones en lengua española, permite reconocer conceptos médicos de un corpus de notas clínicas. Las secciones subsiguientes de esta contribución incluyen primeramente una somera descripción de los trabajos a continuación la descripción de la herramienta y la evaluación de resultados, y para

finalizar con unas breves conclusiones y líneas futuras de las que ya se está elaborando un primer diseño.

2 Trabajos Relativos

La tecnología de información médica tiene como objetivo el procesamiento de notas clínicas con el objetivo de investigar nuevos tratamientos y fármacos. Para ello, la convergencia de diversas disciplinas como las Ciencias de la Computación, Lingüística, la Biomedicina, la Genética, etc. con el objetivo de crear aplicaciones de gestión, consulta y referencia que integren recursos médicos y este objetivo incluye como primera fase o fase de pre-procesamiento, el etiquetado semántico de los documentos.

En el etiquetado semántico de documentos clínicos, el primer paso consiste en la identificación de términos y su mapeo con conceptos. La eficiencia del sistema dependerá de la eficiencia del procesamiento lingüístico y la calidad y cobertura del tesauro utilizado [2]. Con referencia a estos últimos, el uso de meta-tesauros como SNOMED o UMLS, considerados como estándares, aseguran una alta calidad y proporcionan redes semánticas multilingües. En este sentido existen varias aproximaciones e iniciativas que justifican el uso de UMLS [3] y [4], por su amplia cobertura frente a otros recursos terminológicos como GALEN, MeSH o SNOMED [5]. Sin embargo, a pesar de sus claras ventajas, estas terminologías no cubren todos los idiomas, por lo que establecen barreras para los hablantes de lengua no anglosajona y obligan en ocasiones a crear sus propias terminologías [6].

En el dominio biomédico, los registros de pacientes son escritos por expertos humanos lo que puede generar muchos problemas debido al uso excesivo de símbolos que pueden tener varios significados, la construcción parcial de frases, la omisión de lagunas convenciones ortográficas o el uso de términos no normalizados. Por ello, se hace necesaria la incorporación de otro tipo de recursos cono correctores ortográficos, lexicones y diccionarios de siglas para el tratamiento de este tipo de información [7], [8].

3 Reconocedor de Conceptos

El modulo de reconocimiento de conceptos de SNOMED[1], se encuentra dentro de una arquitectura de pre-procesado de textos. Dicha arquitectura ofrece un sistema capaz de extraer información morfo-semántica de un conjunto de notas clínicas de entrada. Como parte de la anterior arquitectura, el modulo de reconocimiento de conceptos, mapea un conjunto de frases procedentes de notas clínicas contra el meta-tesauro SNOMED. Este modulo trata de identificar en las frases de entrada, todos los términos pertenecientes a SNOMED, devolviendo un conjunto de conceptos del meta-tesauro, así como posibles sinónimos o términos relacionados con dichos conceptos.

[1] SNOMED: The systematized nomenclature of medicine. http://www.snomed.org/

El reconocedor de conceptos de SNOMED trabaja de forma análoga a como lo hace la herramienta Metamap[2], que permite reconocer conceptos sobre el sistema médico de lenguaje unificado (UMLS). Más concretamente se trabaja con las tablas en castellano de SNOMED, por lo que el actual reconocedor de conceptos, funcionara con textos en castellano a diferencia de Metamap que lo hace para textos en inglés.

En cuanto al almacenamiento de SNOMED, se han propuesto dos soluciones, indizar la tabla de descripciones con Lucene[3], de manera que se mejore la eficiencia en los accesos a SNOMED por realizarse mediante índices invertidos. Dichos índices han sido construidos de manera que permitan realizar búsquedas por el campo Term de la tabla de descripciones del meta-tesauro. Como segunda opción, se emplea una base de datos MySql[4] desarrollada por la empresa de desarrollo de software ISOCO[5], en la que se incluyen datos de las tres tablas de SNOMED, por lo que si bien la eficiencia seria menor que en el acceso por índices, la semántica recogida en dicha base de datos seria mayor que la de los índices.

4 Evaluación del Reconocedor

La evaluación del reconocedor se realizó sobre los resultados de un corpus de 100 notas clínicas anotadas manualmente por el Gold-Standard (un experto humano) y comparadas con la herramienta objeto del estudio. Además, para esta primera evaluación solo se analizaros las jerarquías "trastornos" y "procedimientos" de SNOMED.

Los indicadores a tener en cuenta fueron:

Umbral de aceptación: Representa el score mínimo que un concepto recuperado debe tener para ser insertado en el árbol de conceptos recuperados.

Numero de conceptos a recuperar: Indica el número de conceptos que se desean recuperar para cada query que se lanza al sistema.

A su vez se probaron varias funciones de evaluación con el fin de encontrar la que mejor se adapta a los datos del fichero Gold-Standard.

Así pues para cada función de evaluación probada, se realizaron seis experimentos, recuperando 1, 2 y 5 conceptos por query y para cada uno de los anteriores con umbral de 0,2 y 0,4.

Por último se tuvieron en cuenta coincidencias totales y laterales y se calcularon los parámetros precisión y recall. En media se obtuvo una precisión de 0.4 y una cobertura de 0.07, valores muy bajos teniendo en cuenta que el reconocedor es capaz de extraer no solo los conceptos sino también los términos relacionados y sus sinónimos.

Analizando el Gold-Standard detectamos la falta de anotación de muchas de las queries de cada nota clínica, lo que lógicamente influye en el resultado. Posteriores trabajos deberían incluir una anotación manual realizada por varios expertos y más

[2] Metamap: http://mmtx.nlm.nih.gov/
[3] Apache Lucene: http://lucene.apache.org/java/docs/index.html
[4] MySql: http://www.mysql.com/
[5] ISOCO: http://www.isoco.com/

exhaustiva

5 Líneas Futuras

Con objeto de refinar el sistema para obtener unos mejores resultados, las líneas futuras incluyen la creación de un repositorio de recursos médicos que, en base a diccionarios terminológicos y/o ontologías consideradas como estándares dentro del ámbito de la medicina, permitan el reconocimiento de términos médicos no incluidos en SNOMED pero vinculados con conceptos propios del meta-tesauro.

Por último, pero no menos importante es ampliar el Gold-Sstandard y el ámbito del corpus de prueba, para obtener una mayor fiabilidad en el contraste de resultados.

References

Ananiadou, S. and McNaught, J. Text Mining for Biology and Biomedicine. Artech House, Inc. (2006).

Vintar, P. Buitelaar, M. Volk, Semantic relations in concept-based cross-language medical information retrieval, in: Proceedings of the Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik (2003).

Volk M., Ripplinger B., Vintar, S., Buitelaar, P., Raileanu, D., Sacaleanu, B. Semantic annotation for concept-based cross-language medical information retrieval. International Journal of Medical Informatics; 67(1): 97-112 (2002).

Jang, H., Song S. K., Myaeng, S. H. Semantic Tagging for Medical Knowledge Tracking. Proceedings of the 28th IEEE EMBS Annual International Conference. New York City, USA, Aug 30-Sept 3 (2006).

Ruch, P., Wagner, J., Bouillon, P., Baud, R., Rassinoux, A.-M., Robert, G. Medtag: Tag-like semantics for medical document indexing. In Proceedings of AMIA'99, p. 35-- 42 (1999).

Lu, W-H., Lin, R., Chan, Y-CH, Chen, K-H. Overcoming Terminology Barrier Using Web Resources for Cross-Language Medical Information Retrieval. AMIA Annu Symp Proc.; 519–523 (2006).

Schuler, K., Kaggal, V., Masanz, J., Ogren, P., Savova, G.. System Evaluation on a Named Entity Corpus from Clinical Notes. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08) (2008).

Ogren, P., Savova, G., Chute, Ch. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08) (2008).

KnowNet: Building a Large Net of Knowledge from the Web

Montse Cuadros[1], Lluís Padró[1], and German Rigau[2]

[1] TALP Research Center, UPC
Barcelona, Spain
cuadros@lsi.upc.edu , padro@lsi.upc.edu
[2] IXA NLP Group, UPV/EHU
Donostia, Spain
german.rigau@ehu.es

Abstract. This paper presents a new fully automatic method for building highly dense and accurate knowledge bases from existing semantic resources. Basically, the method uses a wide-coverage and accurate knowledge-based Word Sense Disambiguation algorithm to assign the most appropriate senses to large sets of topically related words acquired from the web. KnowNet, the resulting knowledge-base which connects large sets of semantically-related concepts is a major step towards the autonomous acquisition of knowledge from raw corpora. In fact, KnowNet is several times larger than any available knowledge resource encoding relations between synsets, and the knowledge KnowNet contains outperform any other resource when is empirically evaluated in a common framework.

1 Building KnowNet

A knowledge net or KnowNet (KN), is an extensible, large and accurate knowledge base, which has been derived by semantically disambiguating small portions of the Topic Signatures(sets of words related to a topic with a weight associated) acquired from the web [1] . Basically, the method uses a robust and accurate knowledge-based Word Sense Disambiguation algorithm to assign the most appropriate senses to the topic words associated to a particular synset. The resulting knowledge-base which connects large sets of topically-related concepts is a major step towards the autonomous acquisition of knowledge from raw text.

We generated four different versions KnowNet by applying the SSI-Dijkstra algorithm to the whole TSWEB (processing the first 5, 10, 15 and 20 words of each of the 35,250 topic signatures). For each TS, we obtained the direct relations from the topic (a word sense) to the disambiguated word senses of the TS (for instance, party#n#1–>federalist#n#1), but also the indirect relations between disambiguated words from the TS (for instance, federalist#n#1–>republican#n#1). Finally, we removed symmetric and repeated relations.

SSI-Dijkstra used only the knowledge present in WordNet and eXtended WordNet which consist of a very large connected graph with 99,635 nodes (synsets) and 636,077 edges (semantic relations).

2 Evaluation framework

In order to empirically establish the relative quality of these new semantic resources, we used the evaluation framework of task 16 of SemEval-2007: Evaluation of wide coverage knowledge resources [2].

In this framework all knowledge resources are evaluated on a common WSD task. In particular, we used the noun-sets of the English Lexical Sample task of Senseval-3 and SemEval-2007 exercises which consists of 20 and 35 nouns respectively. All performances are evaluated on the test data using the fine-grained scoring system provided by the organizers.

Furthermore, trying to be as neutral as possible with respect to the resources studied, we applied systematically the same disambiguation method to all of them. Recall that our main goal is to establish a fair comparison of the knowledge resources rather than providing the best disambiguation technique for a particular knowledge base. All knowledge bases are evaluated as topic signatures.

2.1 Baselines

We have designed a number of baselines in order to establish a complete evaluation framework for comparing the performance of each semantic resource on the English WSD tasks.

RANDOM: For each target word, this method selects a random sense. This baseline can be considered as a lower-bound.

SEMCOR-MFS: This baseline selects the most frequent sense of the target word in SemCor.

WN-MFS: This baseline is obtained by selecting the most frequent sense (the first sense in WN1.6) of the target word. WordNet word-senses were ranked using SemCor and other sense-annotated corpora. Thus, WN-MFS and SemCor-MFS are similar, but not equal.

TRAIN-MFS: This baseline selects the most frequent sense in the training corpus of the target word.

TRAIN: This baseline uses the training corpus to directly build a Topic Signature using TFIDF measure for each word sense and selecting at maximum the first 450 words.

2.2 Other Large-scale Knowledge Resources

In order to measure the relative quality of the new resources, we include in the evaluation a wide range of large-scale knowledge resources connected to WordNet.

WN [3]: This resource uses the different direct relations encoded in WN1.6 and WN2.0. **XWN** [4]: This resource uses the direct relations encoded in eXtended WN.

spBNC [5]: This resource contains 707,618 selectional preferences acquired for subjects and objects from BNC.

spSemCor [6]: This resource contains the selectional preferences acquired for subjects and objects from SemCor.

MCR [7]: This resource integrates the direct relations of WN, XWN and spSemCor.

TSSEM [8]: These Topic Signatures have been constructed using SemCor.

3 Results

KB	P	R	F1	Av. Size
TRAIN	*65.1*	*65.1*	*65.1*	450
TRAIN-MFS	*54.5*	*54.5*	*54.5*	
WN-MFS	*53.0*	*53.0*	*53.0*	
TSSEM	52.5	52.4	52.4	103
SEMCOR-MFS	*49.0*	*49.1*	*49.0*	
MCR[2]	45.1	45.1	45.1	26,429
WN+XWN+KN-20	44.8	44.8	44.8	671
MCR	45.3	43.7	44.5	129
KnowNet-20	44.1	44.1	44.1	610
KnowNet-15	43.9	43.9	43.9	339
spSemCor	43.1	38.7	40.8	56
KnowNet-10	40.1	40.0	40.0	154
(WN+XWN)[2]	38.5	38.0	38.3	5,730
WN+XWN	40.0	34.2	36.8	74
TSWEB	36.1	35.9	36.0	1,721
XWN	38.8	32.5	35.4	69
KnowNet-5	35.0	35.0	35.0	44
WN[3]	35.0	34.7	34.8	503
WN[4]	33.2	33.1	33.2	2,346
WN[2]	33.1	27.5	30.0	105
spBNC	36.3	25.4	29.9	128
WN	44.9	18.4	26.1	14
RANDOM	*19.1*	*19.1*	*19.1*	

KB	P	R	F1	Av. Size
TRAIN	87.6	87.6	87.6	450
TRAIN-MFS	81.2	79.6	80.4	
WN-MFS	66.2	59.9	62.9	
WN+XWN+KN-20	53.0	53.0	53.0	627
(WN+XWN)[2]	54.9	51.1	52.9	5,153
TSWEB	54.8	47.8	51.0	700
KnowNet-20	49.5	46.1	47.7	561
KnowNet-15	47.0	43.5	45.2	308
XWN	50.1	39.8	44.4	96
KnowNet-10	44.0	39.8	41.8	139
WN+XWN	45.4	36.8	40.7	101
SEMCOR-MFS	42.4	38.4	40.3	
MCR	40.2	35.5	37.7	149
TSSEM	35.1	32.7	33.9	428
KnowNet-5	35.5	26.5	30.3	41
MCR[2]	32.4	29.5	30.9	24,896
WN[3]	29.3	26.3	27.7	584
RANDOM	27.4	27.4	27.4	
WN[2]	25.9	27.4	26.6	72
spSemCor	31.4	23.0	26.5	51.0
WN[4]	26.1	23.9	24.9	2,710
WN	36.8	16.1	22.4	13
spBNC	24.4	18.1	20.8	290

Fig. 1. P, R and F1 fine-grained results for the resources evaluated at Senseval-3 and SemEval-07 English Lexical Sample Task, respectively

Table 1 presents ordered by F1 measure, the performance in terms of precision (P), recall (R) and F1 measure (F1, harmonic mean of recall and precision) of each knowledge resource on Senseval-3 and SemEval-07 and the average size of the TS per word-sense. The different KnowNet versions appear marked in bold and the baselines appear in italics.

The different versions of KnowNet consistently obtain better performances as they increase the window size of processed words of TSWEB. As expected, KnowNet-5 obtain the lower results. However, it performs better than WN (and all its extensions) and spBNC. Interestingly, from KnowNet-10, all KnowNet versions surpass the knowledge resources used for their construction (WN, XWN, TSWEB and WN+XWN). In fact, KnowNet-10 also outperforms (WN+XWN)^2 with much more relations per sense. Also interesting is that KnowNet-10 and KnowNet-20 obtain better performance than spSemCor which was derived from annotated corpora. However, KnowNet-20 only performs slightly better than KnowNet-15 while almost doubling the number of relations.

These initial results seem to be very promising. If we do not consider the resources derived from manually sense annotated data (spSemCor, MCR, TSSEM, etc.), KnowNet-10 performs better that any knowledge resource derived by manual or automatic means. In fact, KnowNet-15 and KnowNet-20 outperforms spSemCor which was derived from manually annotated corpora. This is a very interesting result since these KnowNet ver-

sions have been derived only with the knowledge coming from WN and the web (that is, TSWEB), and WN and XWN as a knowledge source for SSI-Dijkstra[3].

Regarding the integration of resources, WN+XWN+KN-20 performs better than MCR and similarly to MCR[2] (having less than 50 times its size). Also interesting is that WN+XWN+KN-20 have better performance than their individual resources, indicating a complementary knowledge. In fact, WN+XWN+KN-20 performs much better than the resources from which it derives (WN, XWN and TSWEB).

4 Conclusions and future research

The knowledge acquisition bottleneck problem is particularly acute for open domain (and also domain specific) semantic processing. The initial results obtained for the different versions of KnowNet seem to be a major step towards the autonomous acquisition of knowledge from raw corpora, since they are several times larger than the available knowledge resources which encode relations between synsets, and the knowledge they contain outperform any other resource when is empirically evaluated in a common framework.

It remains for future research the evaluation of these KnowNet versions in combination with other large-scale semantic resources or in a cross-lingual setting.

References

1. Agirre, E., LopezdeLacalle, O.: Publicly available topic signatures for all wordnet nominal senses. In: Proceedings of LREC, Lisbon, Portugal (2004)
2. Cuadros, M., Rigau, G.: Semeval-2007 task 16: Evaluation of wide coverage knowledge resources. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). (2007)
3. Fellbaum, C., ed.: WordNet. An Electronic Lexical Database. The MIT Press (1998)
4. Mihalcea, R., Moldovan, D.: extended wordnet: Progress report. In: Proceedings of NAACL Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA (2001)
5. McCarthy, D.: Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences. PhD thesis, University of Sussex (2001)
6. Agirre, E., Martinez, D.: Integrating selectional preferences in wordnet. In: Proceedings of GWC, Mysore, India (2002)
7. Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., Vossen, P.: The MEANING Multilingual Central Repository. In: Proceedings of GWC, Brno, Czech Republic (2004)
8. Cuadros, M., Rigau, G., Castillo, M.: Evaluating large-scale knowledge resources across languages. In: Proceedings of RANLP. (2007)

[3] eXtended WordNet only has 17,185 manually labeled senses.

Efficiently managing complex linguistic information

Joseba Alberdi, Xabier Artola, Arantza Díaz, Aitor Soroa

Grupo IXA. Universidad del País Vasco

The IXA team has developed AWA, a general purpose Annotation Web Architecture for representing, storing and accessing the information produced by different linguistic processors. The objective of AWA is to establish a coherent and flexible representation scheme that will be the basis for the exchange and use of linguistic information. In morphologically-rich languages as Basque it is necessary to represent and provide easy access to complex phenomena such as intraword structure, declension, derivation and composition features, constituent discontinuousness (in multiword expressions) and so on. AWA provides a well-suited schema to deal with these phenomena. The annotation model relies on XML technologies for data representation, storage and retrieval. Typed feature structures are used as a representation schema for linguistic analyses. A consistent underlying data model, which captures the structure and relations contained in the information to be manipulated, has been identified and implemented. AWA is integrated into LPAF, a multi-layered Language Processing and Annotation Framework, whose goal is the management and integration of diverse NLP components and resources. Moreover, we introduce EULIA, an annotation tool which exploits and manipulates the data created by the linguistic processors. Two real corpora have been processed and annotated within this framework following AWA.

In this presentation, we will address the problem of efficient storage and retrieval of AWA annotations by means of XML native databases, namely, the Berkeley XML Database. Complex linguistic phenomena can be expressed by the usage of XML-encoded feature structures queriable by means of XPath, XQuery and the like. On the other, the use of a relational database allows an adequate representation of the context to express context-involving queries that return KWIC-like results.

Brief summary of the KYOTO project

German Rigau

IXA NLP Group, UPV/EHU Donostia, Basque Country

1 Introduction

The KYOTO project[1] (ICT-211423) [1] stands for Knowledge Yielding Ontologies for Transition-based Organization. KYOTO is a co-funded by the European Union[2] and by national funding of Taiwan and Japan. The project started in March 2008 and will end in March 2011.

KYOTO will construct a language-independent information system for a specific domain (environment, ecology and biodiversity) anchored in a language-independent ontology that is linked to wordnets in seven languages. For each language, information extraction and identification of lexicalized concepts with ontological entries will be carried out by text miners (Kybots). The mapping of language-specific lexemes to the ontology allows for cross-linguistic identification and translation of equivalent terms. KYOTO is developing a wiki infrastructure for enabling long-range knowledge sharing and transfer across many languages and cultures, addressing the need for global and uniform transition of knowledge beyond the specific domains adressed in the project.

Semantic interoperability in KYOTO is achieved by defining the words and expressions in each language through a shared ontology. The KYOTO ontology will be formal language-independent representation of entities that will be used for inferencing and reasoning. The Wiki environment will help the users to agree on the meaning of the concepts of interest, to share their knowledge and to relate the terms and expressions in their language to this knowledge. This process is guided by automatic acquisition of terms and meanings from the textual documents provided by the users, and through automatic definition extraction techniques which will provide glosses for the acquired terms. The collaborative system will help the users review and edit all acquired information, with a special focus on achieving consensus but also for different views and interpretations across languages and cultures. The users can maintain their own system over time and work towards interoperability by fine-tuning their specifications or adding linguistically and culturally diverse groups.

[1] http://www.kyoto-project.eu/

[2] Co-funded by EU -FP7 ICT Work Programme 2007 under Challenge 4 -Digital libraries and Content, Objective ICT-2007.4.2 (ICT-2007.4.4): Intelligent Content and Semantics (challenge 4.2).

The Wiki environment also generates formal knowledge representations from the conceptual modeling. These representations are not shown to the user directly; computer software will extract detailed information and facts from the document collection in the group. The extraction process will use the agreed-upon ontological patterns and their relation to the words and expressions in each language so that the information can be interpreted in the same way across these languages and cultures. Likewise, the KYOTO system functions as an information and knowledge sharing platform. The system aims to establish cross-linguistic and cross-cultural communication and to support building and maintaining the system by groups of people in a shared domain and area of interest.

Currently, we completed the specification and design phase and we are integrating the first versions of the system components. In the project, we will be working on a restricted set of languages: English, Dutch, Italian, Spanish, Basque, Simplified Mandarin Chinese and Japanese. We also will apply the system to the domain of the environment and specifically to the topic of ecosystem services, a global phenomenon with different linguistic and cultural interpretations. Nevertheless, the system is designed in such a way that it can be used for any language and can be applied to any domain.

2 Problemsadressed

Most domain acquisition systems in the semantic web community model each domain separately and restrict the system to a single language or a limited set of languages. They also require knowledge engineers and language-technology experts to do the modeling. The KYOTO system, by contrast, is specifically designed to build global and cross-cultural consensus about the meaning and interpretation of domain-specific language. It tries furthermore to overcome the technology gap between users and system builders. The users are given control over the engineering task on a level that they can understand and that can be directly implemented for their community. As such it is an open system that can be extended and maintained by the users themselves without requiring skills in knowledge engineering or language technology. The main challenges of the project are:

- Automatic term and concept mining techniques should be of sufficient quality and have sufficient semantic depth, so that the data are useful for experts in the domain who are not trained in knowledge engineering and language technology;
- The users should be able to relate the terms across languages and cultures so as to agree on definitions and share them;
- Terms and concepts should be anchored to generic language databases and ontologies to provide interoperability and sharing to people outside the domain but in the same language communities;
- The term databases and their definitions in the ontology should enable extraction of sufficiently useful information and facts from text repositories for all the related

languages, while at the same time the information should be of sufficient quality and depth;
- The interpretation of the information and facts should be the same across the different languages and cultures;
- The users should be able to specify the information and facts of their interest without having to access the complex underlying knowledge structure through a handful of textual examples from which the system abstracts the relevant underlying patterns;

3 Methods applied in the project

KYOTO is intended to process and harmonize knowledge across language and cultures: to achieve this task it is fundamental to support a layered representation for the output of linguistic and semantic processing. As a consequence, the Kyoto Annotation Format (KAF) has been defined. This representation format provides the basis for all the modules that operate on the language representation in a uniform way. The representation format includes tokenization, segmentation, morpho-syntactic analysis, and semantic analysis of the text.

The semantic processing and management of knowledge in KYOTO is supported by the following ontological and linguistic resources:
- Wordnets in each language: semantic networks relating sets of synonymous words to one another and representing the lexicalization patterns of concepts in a language. The edited terms from the users represent a domain wordnet that is maintained by the wiki-group but that is related to the general language wordnet. The K-LMF, a dialect of ISO LMF, has been developed as a common standardized representational device to enable interoperability and easier integration of different wordnets. Accordingly, K-LMF comes equipped with a set of harmonized linguistic information, or Data Categories, necessary to manage the exchange of data between different individual wordnets and to allow their integration in Kyoto, in view of forming an extensive global grid of lexical resources.
- Generic ontology which represents a shared basic framework for interpreta-tion. Users can add new concepts thus creating the domain extension anchored to the generic ontology.

Relying on the linguistic and ontological resources and on the KAF representation of processed knowledge desribed here, the KYOTO architecture is characterized by the interaction of the following modules:
- Tybots(Term Yielding Robots): Programs that extract term hierarchies from the KAF annotation of text. Term hierarchies are conceptual structures that are the basis for the users to edit their terms and concepts.
- Wikyoto: a wiki environment that is the basic interface to the user through which he can select and define the terms in the domain (the output of the Tybots) and produce ontological representations of concepts. The user also uses Wikyoto to

specify the conceptual (ontological?) patterns that he is interested in for fact extraction. These are stored as Kybot profiles.
- Kybots (Knowledge Yielding Robots): Programs that take conceptual patterns specified by the user (Kybot profiles) and matches these patterns in the text for collecting facts.
- Exploitation tools: various interfaces that will be developed to retrieve the useful information that has been mined by the Kybots from text collections in different languages, including semantically enhanced information retrieval, and the generation of tables and maps. For instance, a user interested in the decline of tropical species, will get a table with instances of such declines accompanied by a map showing the most relevant locations and pointers to the relevant documents.

The project just completed the design and specification phase. Currently, the first prototypes are being developed. An early version will become available in early 2009. The current website includes, among others, papers, deliverables, presentations and demos. For instance, an early baseline retrieval system that allows one to search in over 15,000 documents in 4 different languages: English, Dutch, Italian and Spanish. The current search system does not yet exploit the results of KYOTO but carries out a standards keyword based search. It allows one to fill in a complex environmental issue and to try to compile an answer through retrieval actions. All the searches and their success are logged, as are the answer that is compiled.

References

1. Vossen, P., Agirre, E., Calzolari, N., Fellbaum, C., Hsieh, S.K., Huang, C.R., Isahara, H., Kanzaki, K., Marchetti, A., Monachini, M., Neri, F., Raffaelli, R., Rigau, G., Tesconi, M., VanGent, J.: Kyoto: Asystem for mining, structuring, and distributing knowledge across languages and cultures. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesh, Morroco (2008)