Learning by Reading

Jornadas TIMM 2014
Horacio Rodríguez
TALP (UPC)
Guión

- Para qué
- Introducción
- Qué leer
- Tareas implicadas
- Conclusiones
Para qué

• Limitaciones de los sistemas de PLN que se basan únicamente en procesamiento superficial.
• Necesidad de procesamiento semántico
• Necesidad de conocimiento del mundo (World Knowledge, Common Sense Knowledge)
• Adquisición de este conocimiento.
• Un ejemplo. RTE
RTE

- **Equivalence** (*Paraphrase*): \( expr_1 \leftrightarrow expr_2 \)
- **Entailment**: \( expr_1 \Rightarrow expr_2 \) – more general

- Directional relation between two text fragments: *Text* \((t)\) and *Hypothesis* \((h)\):

\[
\text{\( t \) entails \( h \) (\( t \Rightarrow h \)) if, typically, a human reading \( t \) would infer that \( h \) is most likely true}\
\]
• Discourse Commitment-based Framework
A Revenue Cutter, the ship was named for Harriet Lane, niece of President James Buchanan, who served as Buchanan’s White House hostess.

T1. A Revenue Cutter is a ship.
T2. The ship was named for Harriet Lane.
T3. Harriet Lane was the niece of President James Buchanan.
T4. The niece of Buchanan served as Buchanan’s White House hostess.
T5. A Revenue Cutter was named for Harriet Lane.
T6. A Revenue Cutter was named for the niece of President James Buchanan.
T7. A Revenue Cutter was named for Buchanan’s White House hostess.
T8. A Revenue Cutter was named for a White House hostess.
T9. A Revenue Cutter was named for a hostess.
T10. The niece of a President served as Buchanan’s White House hostess.
T11. The niece of a President served as Buchanan’s hostess.
T12. The niece of a President served as a White House hostess.
T13. The niece of a President served at the White House.
T14. The niece of a President had occupation hostess.
T15. The niece of a President served as a hostess.
T16. Harriet Lane was related to President James Buchanan.
T17. Harriet Lane was the niece of a President.
T18. Harriet Lane was related to a President.
T19. Harriet Lane was related to James Buchanan.
T20. James Buchanan had title of President.
T21. James Buchanan had a White House hostess.
T22. James Buchanan had a hostess.
T23. James Buchanan was associated with the White House.
T24. James Buchanan had a niece.
T25. Harriet Lane served as Buchanan’s White House hostess.
T26. Harriet Lane served as Buchanan’s hostess.
T27. Harriet Lane served as a White House hostess.
T28. Harriet Lane served at the White House.
T29. Harriet Lane had occupation hostess.
T30. Harriet Lane served as a hostess.

Hyp(34): Harriet Lane owned a Revenue Cutter.

Hyp(36): Harriet Lane worked at the White House.
• Algunas de estos compromisos se pueden deducir con facilidad del texto:
  – Harriet Lane es sobrina de James Buchanan
  – James Buchanan (es/fue) presidente
  – Un Revenue Cutter es un barco
  – ...

• Otros no aparecen explícitamente en el texto y deben ser adquiridos como conocimiento del mundo:
  – James Buchanan (es/fue) presidente de USA
  – Los presidentes de USA residen en la Casa Blanca
  – La Casa Blanca está en Washington
  – Azafata es una profesión
  – ...
Nutcracker

Nutcracker, Roma (La Sapienza)

• Johan Bos, 2007
• Components of Nutcracker:
  – The C&C parser for CCG
  – Boxer
  – Vampire, a FOL theorem prover
  – Paradox and Mace, FOL model builders
• Background knowledge
  – WordNet [hyponyms, synonyms]
  – NomLex [nominalisations]
Learning by Reading 9

Nutcracker

• Given a textual entailment pair T/H with text T and hypothesis H:
  – Produce DRSs for T and H
  – Translate these DRSs into FOL
  – **Generate Background Knowledge in FOL**
• Use ATPs to determine the likelihood of entailment
Nutcracker

• ... 
  – Generate Background Knowledge in FOL

• ... 

• MiniWordNets
  • Use hyponym relations from WordNet to build an ontology
  • Do this only for the relevant symbols
  • Convert the ontology into first-order axioms
Nutcracker

• ...  
  – MiniWordNets

• Example text:

  There is no asbestos in our products now. Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes.
∀x(user(x) → person(x))
∀x(worker(x) → person(x))
∀x(researcher(x) → person(x))
\(\forall x (\text{person}(x) \rightarrow \neg \text{risk}(x))\)

\(\forall x (\text{person}(x) \rightarrow \neg \text{cigarette}(x))\)

.........
Nutcracker

• ...  
  – Use ATPs to determine the likelihood of entailment

• Create Background Knowledge for T&H
• Give this to the theorem prover:
  • (BK & T’) \rightarrow H’
• If the theorem prover finds a proof, then we predict that T entails H
Nutcracker

• El problema básico de la aproximación es el uso de BK
  – Los resultados son excelentes en cuanto a precisión pero muy limitados en cuanto a cobertura.
  – WN es claramente insuficiente para representar el BK necesario.
  – Se necesitan otras fuentes para la obtención de BK.
Introducción

• Formas de adquisición
  – Manual
  – A partir de fuentes (semi-)estructuradas preexistentes
    • Wikipedia
    • Linked data
      – Dbpedia
      – Yago
      – Freebase
    • Ontologías de dominio
    • Terminologías
    • Glosarios
  – A partir de textos (LbR)
Introducción

• To build a formal representation of a specific, coherent topic through deep processing of concise texts focused on that topic
• Natural Language Understanding (NLU) + Knowledge Integration (KI)
• Formas varias de representación del conocimiento:
  – RDF triples
    • <CONCEPT, RELATION, CONCEPT>
  – Frames
• Tipo de conocimiento:
  – Episódico
    • Podemos ha obtenido 5 diputados en las elecciones al Parlamento europeo del 25 de mayo de 2014
  – Genérico
    • Los perros ladran
Introducción

- Ejemplo tomado de Barker et al, 2007
  - Dominio CORAZÓN
- Relaciones
  - EVENT-to-ENTITY: agent, donor, instrument, etc.
  - ENTITY-to-ENTITY: has-part, location, material, etc.
  - EVENT-to-EVENT: causes, defeats, enables, etc.
  - EVENT-to-VALUE: rate, duration, manner, etc.
  - ENTITY-to-VALUE: size, color, age, etc.
- Component Library
  - 700 conceptos generales
    - events
      - TRANSFER, COMMUNICATE, ENTER
    - entities
      - PLACE, ORGANISM, CONTAINER.
- Seed concepts
  - Pospositive (10)
    - PUMP, MUSCLE
  - Confusers (20)
    - MUSICAL-INSTRUMENT, SHOE
Introducción

• 1. The heart is a pump that works together with the lungs.
• 2. The heart consists of 4 chambers.
• 3. The upper chambers are called atria, and the lower chambers are called ventricles.
• 4. The right atrium and ventricle receive blood from the body through the veins and then pump the blood to the lungs.
• 5. It pumps blood in 2 ways.
• 6. It pumps blood from the heart to the lungs to pick up oxygen.
• 7. The oxygenated blood returns to the heart.
• 8. It then pumps blood out into the circulatory system of blood vessels that carry blood through the body.
Introducción

• **NLU**
  – Parsing: CONTEX
  – Logical Form Generation: LF toolkit
  – Abductive Expansion and Reformulation: TACITUS

• **KI**
  – Word to Concept Mapping
  – Concept Creation
  – Instance Unification
  – SRL
  – Constraint Assertion
  – Adjective Elaboration
  – KB matching
Introducción

• **Output**
  
  – **Concepts**
    
    • (a subclass of CHAMBER), BLOOD (LIQUID-SUBSTANCE), HEART (PUMPING-DEVICE), LUNG (INTERNAL-ORGAN), OXYGEN (GAS-SUBSTANCE), VEIN (BODY-PART), VENTRICLE (CHAMBER) and VESSEL (BODY-PART).
  
  – **Axioms**
    
    • `<Pumping`
      
      – object Blood
      – destination Lung>`
    
    • `<Heart`
      
      – has-part (exactly 4 Chamber)>
    
    • `<Receive`
      
      – origin Body
      – path Vein
      – recipient Ventricle
      – object Blood>`
Qué leer

- **LbR**
  - Mobius
    - Barker et al, 2007
  - Learning Reader
  - Explanation Agent
    - Forbus, et al, 2009
    - McFate, Forbus, Hinrichs, 2013

- **Commonsense**
  - Eslick, 2006
  - Clark, Harrison, 2009 (DART)
  - Havasi et al, 2009 (ConceptNet)
  - Singh, 2005 (EM-ONE, PHD, Open Mind Common Sense, LifeNet, StoryNet, ShapeNet)

- **EL, SF**

- **Event**
  - Chambers, Jurafsky, 2007, 2008, ...
  - Riloff et al, 2007, ...
  - Filatova, Prager, 2005, 2012
Tareas implicadas

• **Entity Linking (EL)**
• **Slot Filling (SF)**
• **Event Detection**
  – Event enrichment
  – Event Generalization
  – Scenario induction
• **Relation Extraction**
  – Domain restricted
  – Unrestricted
TAC KBP

• **KB**
  – 818,741 nodes
  – From English WP 2008
  – For each node:
    • Facts + Reference document

• **Traks**
  – Cold Start
  – SF
  – EL
  – Event
  – Sentiment
TAC KBP SF

• Tasks
  – English SF
  – Chinese SF
  – Temporal SF
  – Sentiment SF
  – SF Validation
  – Cold start
<table>
<thead>
<tr>
<th><strong>Person Slots</strong></th>
<th><strong>Type</strong></th>
<th><strong>List?</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>per:alternate_names</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:date_of_birth</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>per:age</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>per:country_of_birth</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>per:stateorprovince_of_birth</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>per:city_of_birth</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>per:origin</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:date_of_death</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>per:country_of_death</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>per:stateorprovince_of_death</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>per:city_of_death</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>per:cause_of_death</td>
<td>String</td>
<td></td>
</tr>
<tr>
<td>per:countries_of_residence</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:statesorprovinces_of_residence</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:cities_of_residence</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:schools_attended</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:title</td>
<td>String</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>per:employee_or_member_of</strong></td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:religion</td>
<td>String</td>
<td>Yes</td>
</tr>
<tr>
<td>per:spouse</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:children</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:parents</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:siblings</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:other_family</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>per:charges</td>
<td>String</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## Organization Slots

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>List?</th>
</tr>
</thead>
<tbody>
<tr>
<td>org:alternate_names</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>org:political_religious_affiliation</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>org:top_members_employees</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>org:number_of_employees_members</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>org:members</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>org:member_of</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>org:subsidiaries</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>org:parents</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>org:founded_by</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>org:date_founded</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>org:date_dissolved</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>org:country_of_headquarters</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>org:stateorprovince_of_headquarters</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>org:city_of_headquarters</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>org:shareholders</td>
<td>Name</td>
<td>Yes</td>
</tr>
<tr>
<td>org:website</td>
<td>String</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diagnostic Scores</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------------------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Isv</td>
<td>32.93</td>
<td>38.50</td>
</tr>
<tr>
<td>ARPANI*</td>
<td>29.10</td>
<td>47.83</td>
</tr>
<tr>
<td>RPI-BLENDER</td>
<td>30.62</td>
<td>38.19</td>
</tr>
<tr>
<td>PRIS2013</td>
<td>27.82</td>
<td>35.33</td>
</tr>
<tr>
<td>BIT</td>
<td>22.06</td>
<td>57.86</td>
</tr>
<tr>
<td>Stanford</td>
<td>28.46</td>
<td>32.30</td>
</tr>
<tr>
<td>NYU</td>
<td>17.35</td>
<td>50.70</td>
</tr>
<tr>
<td>UWashington</td>
<td>10.31</td>
<td>59.72</td>
</tr>
<tr>
<td>CMUML</td>
<td>10.63</td>
<td>28.79</td>
</tr>
<tr>
<td>SAFT_KRes</td>
<td>13.43</td>
<td>12.43</td>
</tr>
<tr>
<td>UMass_IESL</td>
<td>18.47</td>
<td>9.43</td>
</tr>
<tr>
<td>utaustin</td>
<td>7.91</td>
<td>21.85</td>
</tr>
<tr>
<td>UNED</td>
<td>9.11</td>
<td>15.08</td>
</tr>
<tr>
<td>Compreno</td>
<td>13.19</td>
<td>8.69</td>
</tr>
<tr>
<td>TALP_UPC</td>
<td>9.67</td>
<td>6.54</td>
</tr>
<tr>
<td>IIRG</td>
<td>3.20</td>
<td>7.38</td>
</tr>
<tr>
<td>SINDI</td>
<td>2.80</td>
<td>7.26</td>
</tr>
<tr>
<td>CohenCMU</td>
<td>3.68</td>
<td>1.69</td>
</tr>
<tr>
<td>LDC</td>
<td>58.35</td>
<td>83.81</td>
</tr>
<tr>
<td>Entity Count</td>
<td>Value Count (Pct)</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td>per:title</td>
<td>33</td>
<td>142 (10.8%)</td>
</tr>
<tr>
<td>org:top_members_employees</td>
<td>41</td>
<td>116 (8.8%)</td>
</tr>
<tr>
<td>org:alternate_names</td>
<td>45</td>
<td>82 (6.2%)</td>
</tr>
<tr>
<td>per:employee_or_member_of</td>
<td>28</td>
<td>72 (5.5%)</td>
</tr>
<tr>
<td>per:children</td>
<td>23</td>
<td>52 (3.9%)</td>
</tr>
<tr>
<td>per:cities_of_residence</td>
<td>30</td>
<td>51 (3.9%)</td>
</tr>
<tr>
<td>per:age</td>
<td>31</td>
<td>51 (3.9%)</td>
</tr>
<tr>
<td>per:date_of_death</td>
<td>36</td>
<td>48 (3.6%)</td>
</tr>
<tr>
<td>per:cause_of_death</td>
<td>33</td>
<td>47 (3.5%)</td>
</tr>
<tr>
<td>per:charges</td>
<td>13</td>
<td>45 (3.4%)</td>
</tr>
<tr>
<td>per:alternate_names</td>
<td>24</td>
<td>45 (3.4%)</td>
</tr>
<tr>
<td>per:countries_of_residence</td>
<td>25</td>
<td>36 (2.7%)</td>
</tr>
<tr>
<td>per:city_of_death</td>
<td>32</td>
<td>35 (2.6%)</td>
</tr>
<tr>
<td>org:country_of_headquarters</td>
<td>34</td>
<td>34 (2.6%)</td>
</tr>
<tr>
<td>org:website</td>
<td>32</td>
<td>32 (2.4%)</td>
</tr>
<tr>
<td>per:origin</td>
<td>28</td>
<td>32 (2.4%)</td>
</tr>
<tr>
<td>per:spouse</td>
<td>23</td>
<td>28 (2.1%)</td>
</tr>
<tr>
<td>per:states_or_provinces_of_residence</td>
<td>23</td>
<td>28 (2.1%)</td>
</tr>
<tr>
<td>per:schools_attended</td>
<td>16</td>
<td>27 (2.0%)</td>
</tr>
<tr>
<td>org:subsidiaries</td>
<td>13</td>
<td>25 (1.9%)</td>
</tr>
<tr>
<td>per:parents</td>
<td>18</td>
<td>25 (1.9%)</td>
</tr>
<tr>
<td>org:city_of_headquarters</td>
<td>23</td>
<td>24 (1.8%)</td>
</tr>
<tr>
<td>org:members</td>
<td>4</td>
<td>22 (1.6%)</td>
</tr>
<tr>
<td>org:founded_by</td>
<td>11</td>
<td>21 (1.6%)</td>
</tr>
<tr>
<td>org:state_or_province_of_headquarters</td>
<td>20</td>
<td>20 (1.5%)</td>
</tr>
</tbody>
</table>
### Alternate Names (query expansion)

<table>
<thead>
<tr>
<th>Query</th>
<th>Name</th>
<th>Alternate names</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF558</td>
<td>Barbara Boxer</td>
<td>1.0 Barbara Levy Boxer, 0.8 Barbara L. Boxer, 0.64 B. L. Boxer, 0.56 B. Boxer, 0.49 Boxer, ...</td>
<td>37</td>
</tr>
<tr>
<td>SF520</td>
<td>Hong Kong Disneyland</td>
<td>1.0 Hong Kong Disneyland, 1.0 HKDL, 0.8 H. Kong Disneyland, 0.8 Hong K. Disneyland, 0.7 Hong Disneyland, 0.7 Kong Disneyland, 0.64 Disneyland, ...</td>
<td>30</td>
</tr>
</tbody>
</table>
TAC KBP SF

• distant supervision
  – A partir de Freebase y Wikipedia
  – Filtrado para la reducción del ruido
  – Wikification, normalización de expresiones temporales,
  – Parsing de dependencias
• QA
• hand-coded rules
• IE
• Bootstrapping
• Aprendizaje no supervisado
TAC KBP EL

• Query
  – Name + reference document

• Desambiguación
  – KB
  – NIL

• Preproceso
  – Query classification (PER, ORG, GPE)
  – Query expansion
  – Proceso lingüístico del documento de Referencia
    • Segmentación en oraciones, tokenización, POS tagging, NERC, Correferencia, WSD, SRL, parsing de dependencias,
Dollars for Death

Anybody &lt;anybod...@canada.com&gt; 

Dieticians

The American Dietetic Association (ADA) has 67,000 members. Their motto is &quot;Everything in moderation.&quot; That includes McDonald's, other fast food restaurants, dairy products, NutraPoison, and sugar-rich soda. Of course, the one concept that they do not limit is donations by various industry groups who delight in seeing the ADA's continuing ....... i4cro(at)earthlink.net
TAC KBP EL

• Query expansion
  – AN generation
    • PER
      – Person name grammars
    • ORG
      – Acronyms, suffixes
    • GPE
      – Gazetteers
        » Geonames, GNIS, DBPEDIA, YAGO, YAGO 2
  – WP hyperlinks
  – Coreference
  – Statistical models
TAC KBP EL

- Candidate generation
  - IR
  - Document semantic analysis and context modeling
  - Collaborative clustering
    - Go beyond single query and single KB entry
    - All entities in the reference document
    - Graph-based clustering
TAC KBP EL

• Candidate ranking
  – VSM (unsupervised similarity between the vector representing the query and the vectors representing the candidates)
  – Supervised classification and ranking
  – Global graph-based ranking
  – Rule based
  – Ranking algorithms
    • SVMRank, MaxEnt, Random Forests, ListNet
TAC KBP EL

- NIL clustering
  - Name string matching
  - Longest mention with within-document coreference
  - HA clustering
  - Global graph-based clustering
  - Linking to larger KB and mapping down
  - Topic modeling, LDA, LSA, WP categories, SUMO nodes
Event

Chambers & Jurafsky

- Narrative Event Chains
  - Partially ordered sets of events centered around a common **Protagonist**.
- Typed Narrative Events

```
<table>
<thead>
<tr>
<th>Events</th>
<th>Roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>A search B</td>
<td>A = Police</td>
</tr>
<tr>
<td>A arrest B</td>
<td>B = Suspect</td>
</tr>
<tr>
<td>B plead C</td>
<td>C = Plea</td>
</tr>
<tr>
<td>D acquit B</td>
<td>D = Jury</td>
</tr>
<tr>
<td>D convict B</td>
<td></td>
</tr>
<tr>
<td>D sentence B</td>
<td></td>
</tr>
</tbody>
</table>
```
Event

Chambers & Jurafsky

L = (X pleads), (X admits), (convicted X), (sentenced X)
O = {(pleads, convicted), (convicted, sentenced), ...}

A detain B  A \in \{police, agent, officer, authorities, troops, official, investigator, ... \}
A confiscate B  B \in \{suspect, government, journalist, \textit{monday}, member, citizen, client, ... \}
A seize B
A raid B
A search B
A arrest B
Event

Chambers & Jurafsky

• A database of Narrative Schemas (LREC 2010)
  • Narrative Schemas (unordered)
    • Various sizes of schemas (6, 8, 10, 12)
    • 1813 base verbs
    • 69 documents
    • 740 events
  
  – Temporal Orderings
    • Pairs of verbs
    • Counts of before and after relations

Authorship

Events: publish sell write translate distribute edit produce read

Role 1: \{ translate-s produce-s sell-s write-s distribute-s publish-s read-s edit-s company author group year microsoft magazine my time firm writer government \}

Role 2: \{ produce-o edit-o sell-o translate-o publish-o read-o write-o distribute-o book report novel article story letter magazine film letters movie show \}
Event

Riloff et al

- Huang & Riloff, 2010
- Semantic tagging
- Bootstrapping approach
  - Seed words
  - Example in veterinary medicine domain
    - [A 14yo doxy]ANIMAL owned by [a reputable breeder]HUMAN is being treated for [IBD]DISEASE with [pred]DRUG.
  - Two steps:
    - Inducing a Contextual Classifier
    - Cross-Category Bootstrapping
Event

Riloff et al
Event

Riloff et al

- Huang & Riloff, 2012
- Event extraction classifiers
  - **TIER** Event Extraction Model
  - Bootstrapping approach
    - 4 classifiers
      - Document level
      - Sentence level
      - Noun phrase level

![Diagram of Event Extraction Process](Image)
Event

Filatova et al

• Domain independent Detection, Extraction and Labeling of Atomic Events

THING: China Airlines Flight 676 from Bali to Taipei crashes
PLACE: Taipei, Taiwan
WHEN: February 16, 1998

TOPIC EXPlication: The flight was from Bali to Taipei. It crashed several yards short of the runway and all 196 on board were believed dead. China Airlines had an already sketchy safety record. This crash also killed many people who lived in the residential neighborhood where the plane hit the ground. Stories on topic include any investigation into the accident, stories about the victims/their families/the survivors. Also on topic are stories about the ramifications for the airline.
• Domain independent Detection, Extraction and Labeling of Atomic Events
  – NE pairs

<table>
<thead>
<tr>
<th>Relation Frequency</th>
<th>First Element</th>
<th>Second Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0212</td>
<td>China Airlines</td>
<td>Taiwan</td>
</tr>
<tr>
<td>0.0191</td>
<td>China Airlines</td>
<td>Taipei</td>
</tr>
<tr>
<td>0.0170</td>
<td>China Airlines</td>
<td>Monday</td>
</tr>
<tr>
<td>0.0170</td>
<td>Taiwan</td>
<td>Monday</td>
</tr>
<tr>
<td>0.0170</td>
<td>Bali</td>
<td>Taipei</td>
</tr>
<tr>
<td>0.0148</td>
<td>Taipei</td>
<td>Taiwan</td>
</tr>
<tr>
<td>0.0148</td>
<td>Bali</td>
<td>Taiwan</td>
</tr>
<tr>
<td>0.0148</td>
<td>Taipei</td>
<td>Monday</td>
</tr>
<tr>
<td>0.0127</td>
<td>Bali</td>
<td>Monday</td>
</tr>
<tr>
<td>0.0127</td>
<td>International Airport</td>
<td>Taiwan</td>
</tr>
</tbody>
</table>
Event

Filatova et al

- Domain independent Detection, Extraction and Labeling of Atomic Events
  - Top connectors

<table>
<thead>
<tr>
<th>Relation</th>
<th>Connector</th>
<th>Connector Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>China Airlines – Taiwan</td>
<td>crashed/VBD</td>
<td>0.0312</td>
</tr>
<tr>
<td></td>
<td>trying/VBG</td>
<td>0.0312</td>
</tr>
<tr>
<td></td>
<td>burst/VBP</td>
<td>0.0267</td>
</tr>
<tr>
<td></td>
<td>land/VB</td>
<td>0.0267</td>
</tr>
<tr>
<td>China Airlines – Taipei</td>
<td>burst/VBP</td>
<td>0.0331</td>
</tr>
<tr>
<td></td>
<td>crashed/VBD</td>
<td>0.0331</td>
</tr>
<tr>
<td></td>
<td>crashed/VBN</td>
<td>0.0198</td>
</tr>
</tbody>
</table>
Event

Filatova et al

- Domain independent Detection, Extraction and Labeling of Atomic Events
  - Output Event

<table>
<thead>
<tr>
<th>First named entity</th>
<th>Second named entity</th>
<th>Connectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>China Airlines</td>
<td>Taiwan; Taipei</td>
<td>crashed/VBD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trying/VBG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>burst/VBP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>land/VB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>killing/VBG</td>
</tr>
</tbody>
</table>
Event

Filatova et al

- Learning occupation related activities for Biographies

<table>
<thead>
<tr>
<th>Artist</th>
<th>Dancer</th>
<th>Physicist</th>
<th>Singer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Born/VBN</td>
<td>Made/VBD</td>
<td>Born/VBN</td>
<td>Said/VBD</td>
</tr>
<tr>
<td>Painted/VBD</td>
<td>Died/VBD</td>
<td>Died/VBD</td>
<td>Born/VBN</td>
</tr>
<tr>
<td>Painted/VBN</td>
<td>Appeared/VBD</td>
<td>Announced/VBD</td>
<td>Died/VBD</td>
</tr>
<tr>
<td>Including/VBG</td>
<td>Been/VBN</td>
<td>Discovered/VBD</td>
<td>Join/VB</td>
</tr>
<tr>
<td>Be/VB</td>
<td>Founded/VBD</td>
<td>Be/VB</td>
<td>Singing/VBG</td>
</tr>
<tr>
<td>Became/VBD</td>
<td>Became/VBD</td>
<td>Including/VBG</td>
<td>Sang/VBD</td>
</tr>
<tr>
<td>Died/VBD</td>
<td>Born/VBN</td>
<td>Became/VBD</td>
<td>Has/VBZ</td>
</tr>
<tr>
<td>Been/VBN</td>
<td>Danced/VBD</td>
<td>Wrote/VBD</td>
<td>Conducting/VBG</td>
</tr>
<tr>
<td>Showed/VBD</td>
<td>Blessed/VBN</td>
<td>Helped/VBD</td>
<td>Made/VBD</td>
</tr>
<tr>
<td>Had/VBD</td>
<td>Perform/VB</td>
<td>Named/VBN</td>
<td>Became/VBD</td>
</tr>
</tbody>
</table>
Conclusiones

• **LbR** es una tarea necesaria y difícil
• Implica la realización de varias subtareas (**SF, EL, Event Detection, Event Enrichment, Scenario Induction, Relation Extraction**) que también son difíciles y despiertan el interés de los investigadores en dos áreas de interés: **NLU y KI**
• Las técnicas empleadas son muy variadas
• Se trata de un campo sumamente activo de investigación