Babelplagiarism: what can BabelNet do for cross-language plagiarism detection?

Roberto Navigli
Joint work with…

Simone Ponzetto

Mirella Lapata

Andrea Moro
Outline

• Motivation: the knowledge acquisition bottleneck
• BabelNet: constructing a large-scale multilingual ontology
• What can BabelNet do for (cross-language) plagiarism detection?
• Conclusions: lessons learned
It’s all about knowledge!

• Intuitively, we all know what knowledge is…
• …and why we need it
• But can we expect computers to know?
• Can’t computers just use, e.g., statistical techniques?
EN: I love chocolate, so I bought a bar in the supermarket.
These are movies in which the music genre, e.g. **rock**, is an important element but not necessarily central to the plot. Examples are Easy Rider (1969), The Graduate (1969), and Saturday Night Fever (1978).
• **EN:** These are movies in which the music genre, e.g. rock, is an important element but not necessarily central to the plot. Examples are Easy Rider (1969), The Graduate (1969), and Saturday Night Fever (1978).

• **IT:** Questi sono i film in cui il genere musicale, ad es roccia, è un elemento importante, ma non necessariamente al centro della trama.
Knowledge of the distribution of underground rock densities can assist in interpreting subsurface geologic structure and rock type.
Knowledge of the distribution of underground rock densities can assist in interpreting subsurface geologic structure and rock type.

La conoscenza della distribuzione di densità di rock underground può aiutare a interpretare in sottosuolo struttura geologica e tipo di roccia.
It’s not that the “big data” approach is bad, it’s just that mere statistics is not enough.
The Knowledge Acquisition Bottleneck

- **Knowledge** is crucial in NLP
  - Word Sense Disambiguation
  - Named Entity Recognition
  - Question Answering
  - *(your favourite NLP task here)*

- However, *providing* knowledge is **difficult** and **costly**

- Various *projects* undertaken to make **lexical knowledge** available in a **machine readable format**
  - WordNet [Fellbaum, 1998]
  - Open Mind Word Expert [Chklovski & Mihalcea, 2002]
  - The WordNetPlus project [Boyd-Graber et al., 2006]
  - OntoNotes [Hovy et al., 2006]
  - EuroWordNet [Vossen, 1998], Multilingual Central Repository [Atserias et al. 2004], …
  - Wikipedia *(collaborative effort)*

Plagiarism detection!
Word Sense Disambiguation in a Nutshell

Spring water can be found at different altitudes

The Richer, The Better

- Highly-interconnected semantic networks have a great impact on knowledge-based WSD even in a fine-grained setting [Navigli & Lapata, IEEE TPAMI 2010]

Babelplagiarism: What can BabelNet do for cross-language plagiarism detection?
Knowledge-based WSD NEEDS (a lot of) Knowledge!

- Knowledge-based approaches have a high potential
  - Lexical knowledge resources only partly available
State of the Art “in a nutshell”

- Knowledge-based approaches have a **higher potential**
  - Lexical knowledge resources **only partly available**
  - Only for **few languages** (e.g. not all 23 EU official languages)
  - **Heterogenous** and with **low coverage**
This is where the ERC (and my project) comes into play

A 5-year ERC Starting Grant (2011-2016) on Multilingual Word Sense Disambiguation (http://lcl.uniroma1.it/multijedi)
**Key Objective 1:** create knowledge for all languages
Multilingual Joint Word Sense Disambiguation (MultiJEDI)

Key Objective 2: use all languages to disambiguate one
BabelNet [Navigli & Ponzetto, ACL 2010; AIJ 2012]

- A wide-coverage multilingual semantic network including both encyclopedic (from Wikipedia) and lexicographic (from WordNet) entries
BabelNet integrates the best of both worlds

**WordNet**

S: (n) balloon (large tough nonrigid bag filled with gas or heated air)

S: (n) balloon (small thin inflatable rubber bag with narrow neck)

**Wikipedia**

Speech balloon

*From Wikipedia, the free encyclopedia*

Speech balloons (also speech bubbles, dialogue balloons or word balloons) are a graphic convention used most commonly in comic books, comic strips and cartoons to allow words (and much less often, pictures) to be understood as representing the speech or thoughts of a given character in the scene. There is often a formal distinction between the balloon that indicates thoughts and the one that indicates words spoken aloud; the former conveys subjective thoughts and is often as a thought balloon.

Balloon (typeface)

*From Wikipedia, the free encyclopedia*

Balloon is a brush script commonly used for signage or display purposes. It was designed in 1939 by Max R. Kaufman, for American Type Founders, in response to Howard Allen Trautman's Cartoon, cut for Elekron Type Foundry in 1936. It had no lowercase letters and was cast in Light, Bold, and Extra.

Balloon (game)

*From Wikipedia, the free encyclopedia*

Balloon, balloon-ball or wind-ball was a game similar to the modern game of volleyball in which a leather ball would be battered by the fist or forearm to prevent it from touching the ground. The game was played in ancient Rome where it was known as folia — the Latin word for a leather bag. Such a ball made of leather was quite heavy and so protection might be used such as a leather glove or...
WordNet [Miller et al., 1990; Fellbaum, 1998]

**WordNet**

- The most widespread *computational lexicon* of English [Miller et al., 1990]
- Concepts are encoded as set of synonyms (*synsets*), e.g.:
  \[
  \{ \text{pop}^2_n, \text{soda}^2_n, \text{soda pop}^1_n, \text{soda water}^2_n, \text{tonic}^2_n \}
  \]
- **Semantic relations** connect pairs of synsets
- For each synset, a *textual definition* (*gloss*) is provided, e.g.:
  “a sweet drink containing carbonated water and flavoring”.
Wikipedia [the online community, 2001-today]

**Wikipedia**

- The largest Web encyclopedia
- Wikipedia pages (Wikipages) encode: **concepts** (**SODA (SOFT DRINK)**) or **named entities** (**FOOD STANDARDS AGENCY**)
- The title of a Wikipage (e.g. **SODA (SOFT DRINK)**) is composed of:
  - lemma (**soda**)
  - possibly, a **sense label** (soft drink vs. sodium carbonate)
- Wikipages contain **hyperlinks** to other Wikipages
- Some Wikipages are redirections to other pages (e.g. **SODA (SODIUM CARBONATE) → SODIUM CARBONATE**)
- Wikipages are manually categorized (e.g. **SOFT DRINKS for SODA**)

Babelplagiarism: What can BabelNet do for cross-language plagiarism detection?  
Roberto Navigli
BabelNet: concepts and semantic relations (1)

- **Concepts** and **relations** in BabelNet are harvested from **WordNet** and **Wikipedia**:
  - **WordNet**
    - synsets ➔ concepts
    - semantic pointers ➔ relations
  - **Wikipedia**
    - pages ➔ concepts
    - hyperlinks ➔ relations
BabelNet: concepts and semantic relations (2)

We encode knowledge as a labeled directed graph

- each vertex represents a Babel synset

- each edge expresses a semantic relation

\[ \text{isa} \quad \text{instance-of} \quad \text{part-of} \quad \text{UNSPEC} \]
BabelNet: objectives

1. Provide a unified resource
   - By establishing an automated mapping between Wikipedia pages and WordNet senses

2. Enable multilinguality
   - By collecting the lexicalizations of concepts in different languages using:
     a) Wikipedia interlanguage links
     b) Statistical Machine Translation
Bunescu & Pasca [2006] and Mihalcea [2007] used Wikipedia pages as word senses.

Mihalcea [2007] manually mapped Wikipedia pages to WordNet senses and performs lexical-sample WSD.

**Our contribution:** we fully automatize the mapping between Wikipedia and WordNet.

- We select the most likely WordNet sense $s$ of a wikipedia page $w$:

$$
\mu(w) = \begin{cases} 
  s \in Senses_{WN}(w) & \text{if a link can be established}, \\
  \epsilon & \text{otherwise}.
\end{cases}
$$
An example of mapping
Creation of the Wikipedia disambiguation contexts

- **Wikipedia**: given a page (e.g. BALLOON (AIRCRAFT))
  - sense labels
  - links
  - categories

aircraft
wind, gas, helium, ...
technology
Building BabelNet: Mapping Wikipedia to WordNet (2)

\[ \mu(w) = \arg\max_{s \in Senses_{WN}(w)} p(s \mid w) = \arg\max_s \frac{p(s, w)}{p(w)} = \arg\max_s p(s, w) \]

- Given a Wikipedia page \( w \) and its disambiguation context \( \text{ctx}(w) \):
  - For each WordNet sense \( s \) of \( w \), calculate score\((s, w)\) as follows:

\[
\text{score}(s, w) = \sum_{cw \in \text{ctx}(w)} \sum_{s' \in Senses_{WN}(cw)} \sum_{p \in \text{paths}_{WN}(s, s')} e^{-(\text{length}(p)-1)}
\]
Building BabelNet: Translating Babel synsets

1. Exploiting Wikipedia interlanguage links

1. Exploiting Wikipedia interlanguage links
Building BabelNet: Translating Babel synsets

2. Filling the lexical translation gaps using a Machine Translation system to translate the English lexicalizations of a concept

- On August 27, 1783 in Paris, Franklin witnessed the world's first hydrogen flight.
- Le 27 Août, 1783 à Paris, Franklin vu le premier vol en ballon d'hydrogène.
Building BabelNet: Translating Babel synsets

2. **Filling the lexical translation gaps using a Machine Translation system to translate** the English lexicalizations of a concept

- For each word sense $s$, we translate:
  - sentences from **SemCor** (a corpus annotated with WordNet senses) which contain $s$
  - sentences from **Wikipedia** linked to the Wikipage of $s$
- The **most frequent translation of $s$ is selected for each target language**

![Diagram showing translations between languages: English, German, Italian, French, Spanish, and Catalan.](image-url)
BabelNet: an encyclopedic dictionary!

- Available online: http://babelnet.org

For research purposes...
Anatomy of BabelNet

• **6 languages covered (moving to 40+)**
• **More than 3 million Babel synsets (i.e. concepts and NE)**
• **More than 26 million word senses:**

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Catalan</th>
<th>French</th>
<th>German</th>
<th>Italian</th>
<th>Spanish</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>English WordNet</td>
<td>206,978</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>206,978</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>2,955,552</td>
<td>123,101</td>
<td>524,897</td>
<td>506,892</td>
<td>404,153</td>
<td>349,375</td>
<td>4,863,970</td>
</tr>
<tr>
<td>- pages</td>
<td>3,388,049</td>
<td>105,147</td>
<td>617,379</td>
<td>456,977</td>
<td>217,963</td>
<td>404,009</td>
<td>5,189,524</td>
</tr>
<tr>
<td>- redirections</td>
<td>-</td>
<td>3,445,470</td>
<td>2,844,751</td>
<td>2,841,916</td>
<td>3,046,325</td>
<td>3,083,427</td>
<td>15,261,889</td>
</tr>
<tr>
<td>WordNet</td>
<td>-</td>
<td>97,876</td>
<td>98,081</td>
<td>97,677</td>
<td>98,475</td>
<td>98,092</td>
<td>490,196</td>
</tr>
<tr>
<td>- monosemous</td>
<td>-</td>
<td>6,852</td>
<td>6,855</td>
<td>6,850</td>
<td>6,856</td>
<td>6,855</td>
<td>34,268</td>
</tr>
<tr>
<td>Total</td>
<td>6,550,579</td>
<td>3,778,446</td>
<td>4,091,963</td>
<td>3,910,307</td>
<td>3,773,772</td>
<td>3,941,758</td>
<td>26,046,825</td>
</tr>
</tbody>
</table>

• **About 70 million lexico-semantic relations:**

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Catalan</th>
<th>French</th>
<th>German</th>
<th>Italian</th>
<th>Spanish</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet</td>
<td>364,552</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>364,552</td>
</tr>
<tr>
<td>WordNet glosses</td>
<td>617,785</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>617,785</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>50,104,898</td>
<td>971,379</td>
<td>5,594,590</td>
<td>5,931,099</td>
<td>3,598,733</td>
<td>3,397,754</td>
<td>69,598,453</td>
</tr>
<tr>
<td>Total</td>
<td>51,087,235</td>
<td>971,379</td>
<td>5,594,590</td>
<td>5,931,099</td>
<td>3,598,733</td>
<td>3,397,754</td>
<td>70,580,790</td>
</tr>
</tbody>
</table>
Evaluation of the Wikipedia-WordNet mapping

- Test set of 1,000 Wikipages manually mapped to the corresponding WordNet sense, if available

<table>
<thead>
<tr>
<th>Mapping method</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BoW</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>taxonomic</td>
<td>89.7</td>
<td>47.8</td>
<td>62.3</td>
<td>72.6</td>
</tr>
<tr>
<td>gloss</td>
<td>87.6</td>
<td>51.8</td>
<td>65.1</td>
<td>74.0</td>
</tr>
<tr>
<td>taxonomic + gloss</td>
<td>87.5</td>
<td>65.6</td>
<td>75.0</td>
<td>80.9</td>
</tr>
</tbody>
</table>

| Graph                   |      |      |      |      |
| taxonomic relations     |      |      |      |      |
| max depth @ 2           | 87.2 | 60.8 | 71.6 | 77.9 |
| max depth @ 3           | 81.6 | 65.0 | 72.4 | 78.7 |
| max depth @ 4           | 78.3 | 69.5 | 73.6 | 79.4 |

| gloss relations         |      |      |      |      |
| max depth @ 2           | 80.5 | 60.6 | 69.1 | 77.0 |
| max depth @ 3           | 77.5 | 65.2 | 70.9 | 78.2 |
| max depth @ 4           | 72.4 | 67.1 | 69.6 | 78.0 |

| taxonomic + gloss relations |      |      |      |      |
| max depth @ 2             | 81.2 | 74.6 | **77.7** | **82.7** |
| max depth @ 3             | 72.8 | **77.4** | 75.1 | 80.1 |
| max depth @ 4             | 64.3 | 76.2 | 69.8 | 75.0 |

| MFS baseline             | 25.4 | 49.2 | 33.5 | 25.4 |
| Random baseline          | 24.2 | 46.9 | 31.9 | 24.2 |
### Evaluation of BabelNet against gold standard resources

#### Coverage

<table>
<thead>
<tr>
<th>Language</th>
<th>Resource</th>
<th>Method</th>
<th>SENSES</th>
<th>SYNSETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>WIKI</td>
<td>Links</td>
<td>39.6</td>
<td>50.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transl.</td>
<td>42.6</td>
<td>58.2</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>Transl.</td>
<td>21.0</td>
<td>28.6</td>
</tr>
<tr>
<td></td>
<td>BABELNET</td>
<td>All</td>
<td>57.6</td>
<td>73.4</td>
</tr>
<tr>
<td>Spanish</td>
<td>WIKI</td>
<td>Links</td>
<td>34.4</td>
<td>40.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transl.</td>
<td>47.9</td>
<td>56.1</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>Transl.</td>
<td>25.2</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>BABELNET</td>
<td>All</td>
<td>66.4</td>
<td>76.6</td>
</tr>
<tr>
<td>Catalan</td>
<td>WIKI</td>
<td>Links</td>
<td>20.3</td>
<td>25.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transl.</td>
<td>46.9</td>
<td>54.1</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>Transl.</td>
<td>25.0</td>
<td>29.6</td>
</tr>
<tr>
<td></td>
<td>BABELNET</td>
<td>All</td>
<td>64.0</td>
<td>73.3</td>
</tr>
<tr>
<td>Italian</td>
<td>WIKI</td>
<td>Links</td>
<td>28.1</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transl.</td>
<td>39.9</td>
<td>58.0</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>Transl.</td>
<td>19.7</td>
<td>28.7</td>
</tr>
<tr>
<td></td>
<td>BABELNET</td>
<td>All</td>
<td>52.9</td>
<td>73.7</td>
</tr>
<tr>
<td>French</td>
<td>WIKI</td>
<td>Links</td>
<td>70.0</td>
<td>72.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transl.</td>
<td>69.6</td>
<td>79.6</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>Transl.</td>
<td>16.3</td>
<td>19.4</td>
</tr>
<tr>
<td></td>
<td>BABELNET</td>
<td>All</td>
<td>86.0</td>
<td>92.9</td>
</tr>
</tbody>
</table>
Evaluation of BabelNet against gold standard resources

Extra-coverage

(a) word senses

(b) synsets
# Coarse-grained Word Sense Disambiguation with BabelNet

<table>
<thead>
<tr>
<th>Resource</th>
<th>Algorithm</th>
<th>Nouns only P/R/F₁</th>
<th>All words P/R/F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet</td>
<td>Degree</td>
<td>80.1</td>
<td>79.7</td>
</tr>
<tr>
<td></td>
<td>PLenght</td>
<td>80.2</td>
<td>79.8</td>
</tr>
<tr>
<td></td>
<td>SProbability</td>
<td>79.8</td>
<td>79.3</td>
</tr>
<tr>
<td></td>
<td>PageRank</td>
<td>79.9</td>
<td>79.4</td>
</tr>
<tr>
<td>BabelNet</td>
<td>Degree</td>
<td>84.7</td>
<td>82.3</td>
</tr>
<tr>
<td></td>
<td>PLenght</td>
<td>85.4</td>
<td>82.7</td>
</tr>
<tr>
<td></td>
<td>SProbability</td>
<td>84.6</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>PageRank</td>
<td>82.3</td>
<td>80.1</td>
</tr>
<tr>
<td></td>
<td>SUSSX-FR</td>
<td>81.1</td>
<td>77.0</td>
</tr>
<tr>
<td></td>
<td>TreeMatch</td>
<td>N/A</td>
<td>73.6</td>
</tr>
<tr>
<td></td>
<td>NUS-PT</td>
<td>82.3</td>
<td>82.5</td>
</tr>
<tr>
<td></td>
<td>SSI</td>
<td>84.1</td>
<td>83.2</td>
</tr>
<tr>
<td></td>
<td>MFS BL</td>
<td>77.4</td>
<td>78.9</td>
</tr>
<tr>
<td></td>
<td>Random BL</td>
<td>63.5</td>
<td>62.7</td>
</tr>
</tbody>
</table>
Main alternatives to BabelNet

• **WikiNet** [Nastase et al., 2011]
  – a multilingual semantic network built from Wikipedia and including semantic relations between Wikipedia entities collected from the category network, infoboxes and article bodies

• **Universal WordNet** [de Melo & Weikum, 2009]
  – bootstrapped from WordNet and built by collecting evidence extracted from existing wordnets, translation dictionaries, and parallel corpora

• **MENTA** [de Melo & Weikum, 2010]
  – multilingual taxonomy containing 5.4 million entities, also built from WordNet and Wikipedia using a number of heuristics

<table>
<thead>
<tr>
<th>Resource</th>
<th>Lemmas</th>
<th>Concepts</th>
<th>Word senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>UWN</td>
<td>822,212</td>
<td>117,659</td>
<td>1,595,763</td>
</tr>
<tr>
<td>MENTA</td>
<td>837,627</td>
<td>82,115</td>
<td>845,210</td>
</tr>
<tr>
<td></td>
<td>upper-level</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>full</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WikiNet</td>
<td>11,721,594</td>
<td>3,707,718</td>
<td>14,200,945</td>
</tr>
<tr>
<td>BabelNet</td>
<td>23,936,234</td>
<td>3,032,406</td>
<td>26,045,741</td>
</tr>
</tbody>
</table>
We developed the BabelNet API for effectively accessing multilingual semantic networks such as BabelNet
  – A Java API based on Apache Lucene
  – Available at: http://babelnet.org

We created a Web application for visualizing and exploring semantic networks
  – Based on Cytoscape Web, a state-of-the-art visualization software

Available at: http://lcl.uniroma1.it/bnxplorer
The BabelNet API

BabelNet bn = BabelNet.getInstance();
System.out.println("SYNSETS WITH English word: " + "bank");
List<BabelSynset> synsets = bn.getSynsets(Language.EN, "bank");
for (BabelSynset synset : synsets)
{
    System.out.println(" =>\" + synset.getId() + ") SOURCE: " + synset.getSource() +
    "; WN SYSET: " + synset.getWordNetOffsets() + ";\" +
    " MAIN LEMMA: " + synset.getMainLemma() + ";\n SENSES (German): { ");
    for (BabelSense sense : synset.getGenses(Language.DE))
        System.out.println(sense.toString() + ");
    System.out.println(" \-----\" + synset.getKeyset());
    Map<IPointer, List<BabelSynset>> relatedSynsets = synset.getRelatedSynsets();
    for (IPointer relationType : relatedSynsets.keySet())
    {
        List<BabelSynset> relationSynsets = relatedSynsets.get(relationType);
        for (BabelSynset relationSynset : relationSynsets)
        {
            System.out.println(" EDGE \" + relationType.getSymbol() +
                " \" + relationSynset.getId() +
                " \" + relationSynset.toString(Language.EN));
        }
    }
    System.out.println(" \-----\");
}
BabelNetXplorer: semantic network exploration

- Type a (possibly ambiguous) word in any language:
BabelNetXplorer: semantic network exploration

• Click a Babel sense of the input word:

Babelplagiarism: What can BabelNet do for cross-language plagiarism detection?
Roberto Navigli

21/09/2012 44
BabelNetXplorer: semantic network exploration

• Expand the graph by clicking on a node:

Expand with the neighbours of the selected node
BabelNetXplorer: semantic network exploration

• Expand the graph by clicking on a node:
BabelNetXplorer: search for connecting paths

• Search the graph for connecting paths:
public static void acl12demo() throws IOException {
    List<Word> sentence =
        Arrays.asList(new Word[] {
            new Word("bank", 'n', Language.EN),
            new Word("bonus", 'n', Language.EN),
            new Word("pay", 'v', Language.EN),
            new Word("stock", 'n', Language.EN)
        });
    disambiguate(sentence, KnowledgeBase.BABELNET, KnowledgeGraphScorer.DEGREE);
}

public static void disambiguate(Collection<Word> words,
                                  KnowledgeBase kb,
                                  KnowledgeGraphScorer scorer) throws IOException {
    KnowledgeGraphFactory factory = KnowledgeGraphFactory.getInstance(kb);
    KnowledgeGraph kGraph = factory.getKnowledgeGraph(words);
    Map<String, Double> scores = scorer.score(kGraph);
    for (String concept : scores.keySet()) {
        double score = scores.get(concept);
        for (Word word : kGraph.wordsForConcept(concept))
            word.addLabel(concept, score);
    }
    for (Word word : words) {
        System.out.println("\n\t" + word.getWord() + " -- ID " +
            word.getId() + " -- SENSE DISTRIBUTION: ");
        for (ScoredItem<String> label : word.getLabels())
            System.out.println("\t [" + label.getItem() + "]: " +
                Strings.format(label.getScore()));
    }
}

Target words can even be in mixed languages!
Create a disambiguation graph for the target words
And disambiguate in 1 line!
Coming soon to your screens: BabelNet 1.1!

BabelNet 1.1
A very large multilingual ontology

search  disambiguate

publications  download

means: 40 languages + more accurate mappings and translations!

Roberto Navigli
Now… why am I saying all this to YOU?!

He is trying to steal important secrets from us…
Plagiarism detection: the state of the art

- Stemming, stopword removal, chunking into passages, keyphrase extraction, n-grams, query formulation, search control, etc.
So, what can we do? [Examples from Vila et al. 2011]

- Same polarity substitutions:
  Google bought YouTube  Google purchased YouTube
So, what can we do?

• Opposite polarity substitutions:  
  Google bought YouTube  
  YouTube was sold to Google
So, what can we do?

- **Deletion:**
  I like eating chocolate  
  I like chocolate
So, what can we do?

- **Semantics based changes:**

  Bill flew across the ocean  Bill crossed the ocean by plane
Remember? BabelNet is multilingual!

• So one sentence can be in **English**, one in **Italian**

Paolo is eating **Parmesan**  
Paolo sta mangiando il parmigiano

• However, note that only **nominal concepts** and **Named Entities** are multilingual!
  – verbs, adjectives and adverbs only in English
Conclusions

- **Statistics** alone is not enough!
- We provide a (hopefully useful) tool for **multilingual lexical semantics**
- This includes **cross-language plagiarism detection**!
- You just have to **download BabelNet** and start **coding**!
What comes next…

- Plenty of work to do!
- **BabelNet:**
  - Increasing the **accuracy** of BabelNet (e.g. game with a purpose)
  - Integrate **more knowledge** (Wikipedia categories, Wiktionary, adjectives, verbs, etc.)
  - **Labeling** relatedness relations (see WiSeNet [Moro & Navigli, CIKM 2012])
  - More languages (40+)
- Much more!
Thanks or…

merci
(grazie)
Roberto Navigli
Linguistic Computing Laboratory
http://lcl.uniroma1.it

Joint work with: Simone Ponzetto; +Mirella Lapata, +Andrea Moro