Cross-lingual named entity disambiguation for concept translation

Tadej Štajner

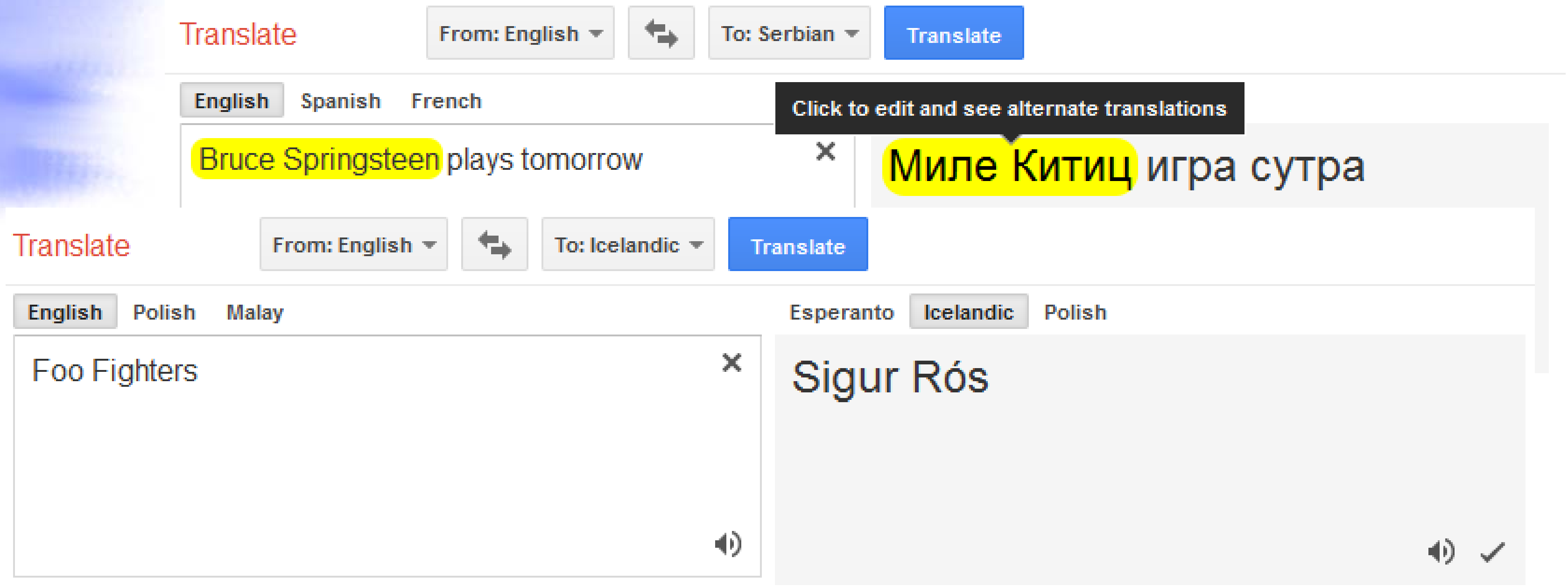
Jožef Stefan Institute

15 March 2012, Luxembourg

W3C Workshop: The Multilingual Web – The Way Ahead

Motivation

- Translating proper names
- … can be problematic for statistical MT systems



Click to edit and see alternate translations

Bruce Springsteen plays tomorrow

Миле Китиц игра сутра

Foo Fighters

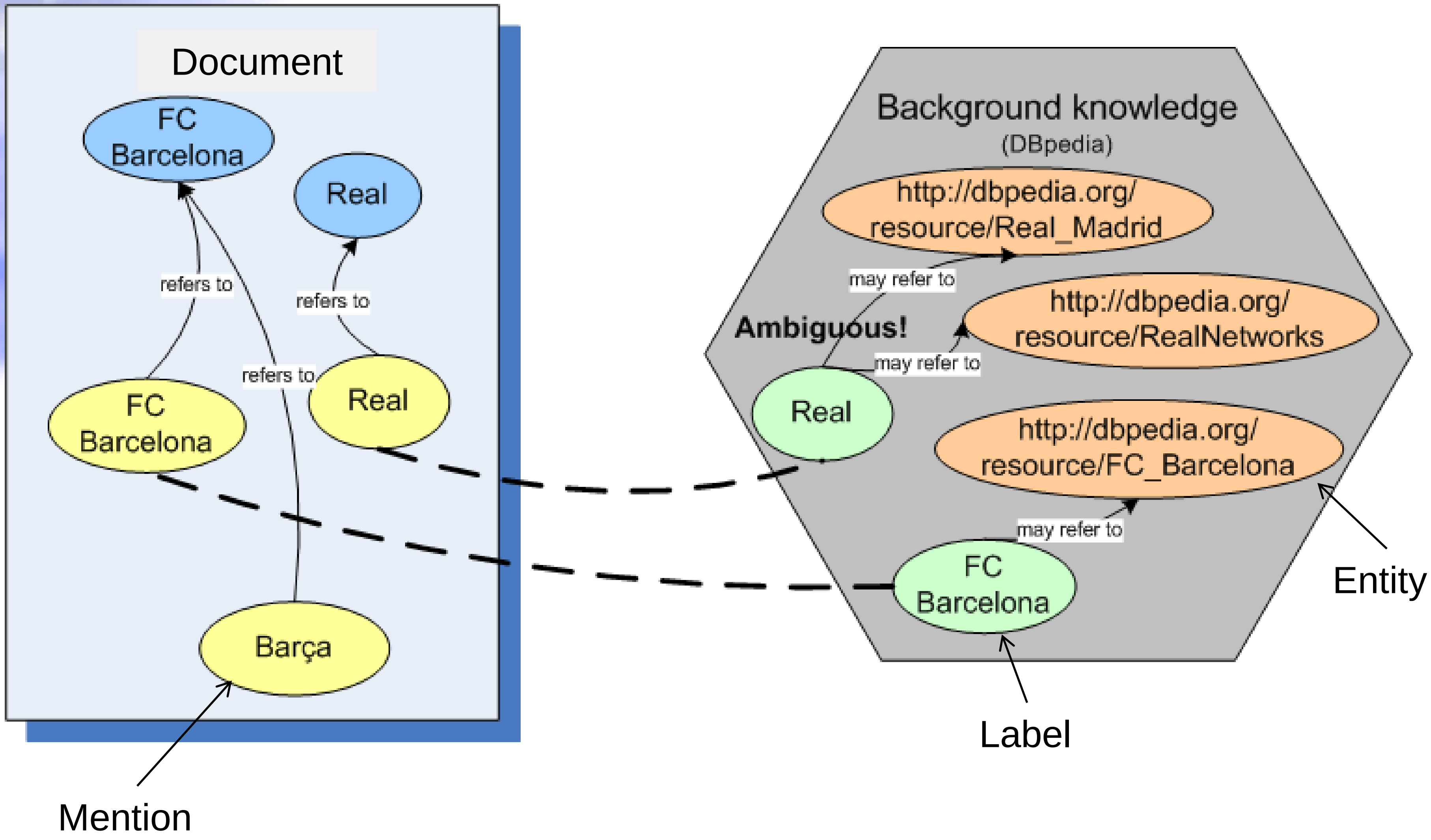
Sigur Rós

- HTML5 `translate` attribute helps, but someone still needs to do the actual mark-up

Motivation (2)

- Depends on source and target language:
  - There are specific rules to translate (or transliterate) particular proper names or concepts
  - Sometimes, they should not even be translated
- Possible solution: **figure out what entity is actually being mentioned and see if any existing translated expression exists for that entity:**
  - Using a background knowledge base
  - Translates the problem into named entity disambiguation

Named entity disambiguation



Knowledge bases

- Doing this requires good coverage of entities in the KB
- The usual choice is DBPedia
- Works well for the bigger languages (en)
  - What about languages with less coverage?
  - as of Jan 2012, English has 3.9M articles, Slovene has 132k*

*http://stats.wikimedia.org/EN/TablesArticlesTotal.htm

Cross-lingual named entity disambiguation

- What if the input document and the knowledge base are in different languages?
  - … there is no knowledge base for a particular language
  - ... the proper knowledge base is too sparse
- Can we share these knowledge bases **across languages**, given that they have different coverage?

Important ranking features

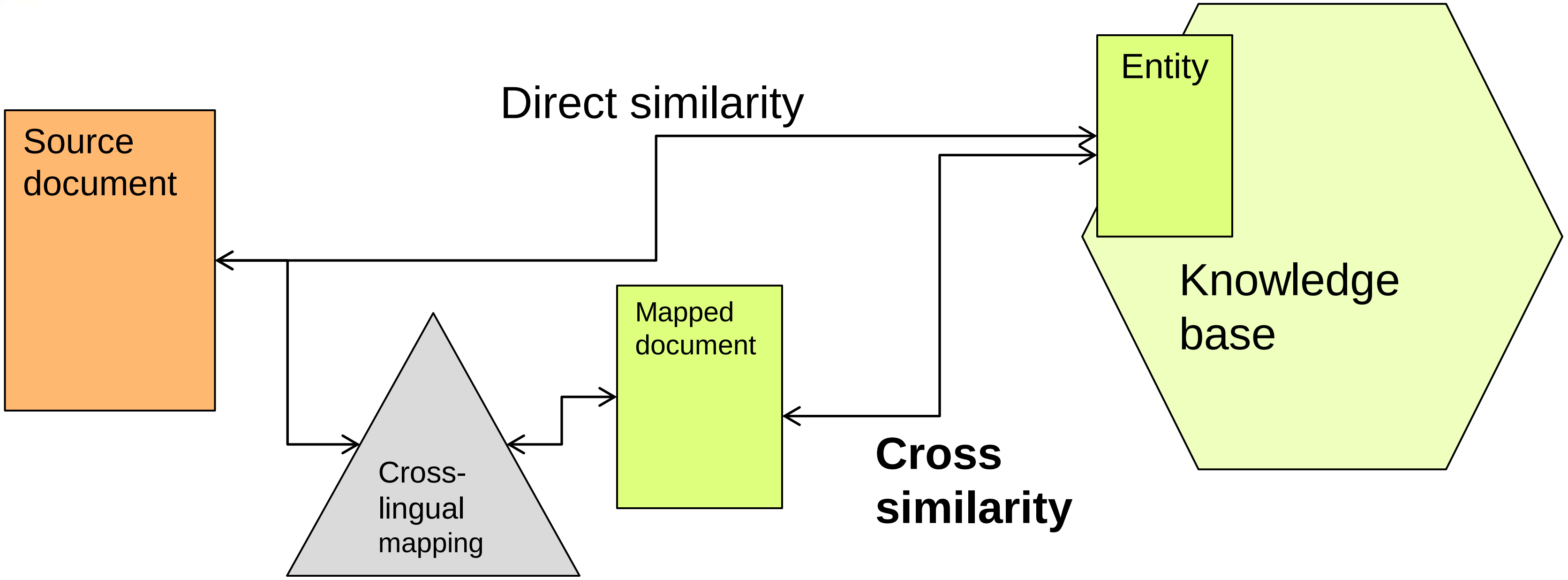
- Mention popularity – `P(entity|mention)`
  - ”Kashmir” .. Kashmir_(song) = 0.05
  - ”Kashmir” … Kashmir_(region) = 0.91
  - **Captures the most likely entity behind the mention**
- Context similarity - `sim(ctx(mention), ctx(entity))`
  - Context of a mention: surrounding sentences
  - Context of an entity: the description of the entity
  - **Captures the entity that best fits the lexical context**
- Coherence
  - Entities that appear together tend to be related to one another
  - Usually solved by a greedy graph pruning algorithm
  - **Collectively captures the entities that make sense appearing together**

What breaks when going cross-lingual?

- Gathering candidate entities for a label
  - Only works reliably for proper names, and even that only when there’s no transliteration or the KB has the concept name in a local language
- Mention popularity
  - (same problem)
- Context similarity
  - Similarity operates in vector space, treating the distinct words as dimensions.
  - Across different languages the words don’t line up, so the similarity is almost meaningless!

Cross-lingual context similarity

- Instead of just directly computing similarity, map the input document into the target language via a mapping, and compute similarity in that space.

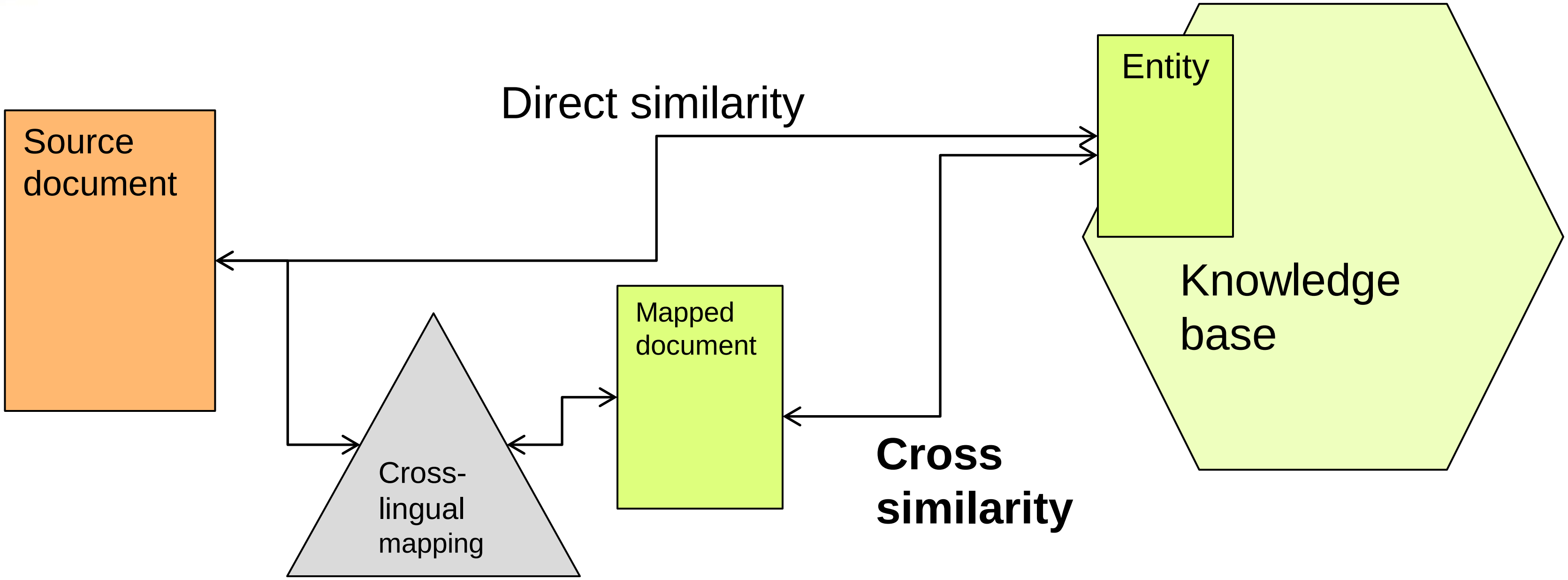


How do we obtain the mapping?

- We train it via a parallel (or comparable) corp
  - Not statistical MT – just providing a linear mapping from one language space to another, which is an easier problem to solve
  - CLIR technique: Canonical Correlation Analysis
  - Our implementation: EuroParl

Potential issues

- If the mapping is weak because of low domain overlap, back off to direct similarity



Future work

- Re-use language and semantic resources to improve performance on NLP tasks across different languages
  - FP7 - XLike
- Lower the barrier for using this technology for enriching content within a CMS
  - standardization work in the W3C Multilingual Web – LT WG

How to make this technology useful?

- Use these annotations within HTML
- Transparent to:
  - Normal CMS operation
  - Web browser rendering
- Readable to:
  - Localization workflow (terminology management - ITS)
  - Downstream NLP processing (OLiA, NIF)
  - Metadata crawlers (knowledge management)
  - Training of MT systems

Demo

- Example (RDFa Lite)
  - enrycher.ijs.si