Multi-lingual Concept Extraction with Linked Data and Human-in-the-Loop

Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, Steve Welch

IBM Research
Motivation

sushirritto

Animal style fries
Motivation

- extract information from a novel corpus
- what are the relevant concepts in the domain?
- limited domain and language knowledge
- IDEA: combine statistical techniques with user-in-the-loop
Domain Learning Assistant

- Start with a small number of seeds (1)
- Get suggestions of new surface forms
- The user accept/reject
Finding concept candidates

The safety and efficacy of **filgrastim** are similar in adults and children receiving cytotoxic chemotherapy

La eficacia y la seguridad del **filgrastim** son similares en los adultos y en los niños tratados con quimioterapia citotóxica

La sicurezza e l’efficacia del **filgrastim** sono simili negli adulti e nei bambini sottoposti a chemioterapia citotossica

Die Wirksamkeit und Unbedenklichkeit von **Filgrastim** ist bei Erwachsenen und bei Kindern, die eine zytotoxische Chemotherapie erhalten, vergleichbar
Plasma elimination half-life of oral pravastatin is 1.5 to 2 hours.

L’emivita plasmatica di eliminazione del pravastatin orale è compresa tra un’ora e mezzo e due ore.
Finding **concept candidates**

Candidates: \{eggs, flour\}

“mix *eggs* and *flour*” $\rightarrow$ mix <candidate> and <candidate>

mix <candidate> and <candidate> $\rightarrow$ “mix *sugar* and *butter*”

Candidates: \{eggs, flour, sugar, butter\}

“melt the *butter*” $\rightarrow$ melt the <candidate>

...
Finding concept candidates

Candidates: \{uova, farina\}

“amalgamare uova e farina” \rightarrow amalgamare <candidate> e <candidate>

amalgamare <candidate> e <candidate> \rightarrow “amalgamare zucchero e burro”

Candidates: \{uova, farina, zucchero, burro\}

“sciogliere il burro” \rightarrow sciogliere il <candidate>

...
Multi-lingual experiment

**HYPOTHESIS**: same behavior, regardless of the language

- we start with very **few seeds** (one could be sufficient) for each language
- we extract context patterns and use them to **generate new candidates**
- we ask to **user** to **accept/reject** the candidates
- we repeat for a fixed number of iterations in all languages
Multi-lingual experiment: Drug Discovery

- **DATA**: parallel corpus from the European Medicines Agency (EMEA)
  - documents related to **medicinal products**
  - translations into 22 official languages of the European Union
  - 1,500 documents for most of the languages
  - we used 4 languages (**en, es, it, de**)

- **TASK**: build a lexicon of clinical drugs

**user-in-the-loop** simulated by constructing a Gold Standard (GS) of drugs names extracted from Linked Open Data (we used **DBpedia** http://dbpedia.org)
Drug Discovery: One seed

- **initial seeds**: single seed
  - One drug name which appears in each corpus (e.g. “irbesartan”)
- 20 iterations
- learning curves for all languages are comparable

Discovery growth for glimpse for English (en), Italian (it), Spanish (es) and German (de). Average correlation amongst all languages $r = 0.998$. 
Drug Discovery: Linked Data seeds

- **initial seeds**: 20% of available Linked Data (DBpedia)
  - 5-fold validation (randomly selected 20%, same drugs for all languages)
  - choice of initial seeds **does not impacts** the results

Discovery growth with 5-fold cross validation on the EMEA dataset using DBpedia as seeds. Each plot shows the discovery growth for each of the randomly generated 5 folds and reports the Pearson correlation (r) amongst them.

(a) EMEA English \( r = 0.991 \)
(b) EMEA German \( r = 0.995 \)
(c) EMEA Spanish \( r = 0.994 \)
(d) EMEA Italian \( r = 0.996 \)
Drug Discovery: benefit of Linked Data

- glimpse \(\rightarrow\) one manually provided seed
- glimpseLD \(\rightarrow\) Linked Data seeds
- in 10 iterations glimpseLD can cover the same lexicon that would take more than 20 iterations with glimpse

Human-in-the-loop experiment with a subject matter expert (physician)

(a) Discovery growth for glimpseLD.

(b) Recall for glimpse vs glimpseLD.
Multi-lingual experiment: Colors

- **DATA**: Twitter stream 1st-14th of January 2016 – lang: En, De, Es, It
  - contain at least one mention of a color
    - gold standard lists of colors from Wikidata and Dbpedia
  - balance datasets size in different languages
    - 155, 828 tweets per language
- **TASK**: expand the lexicon of colors
- **user-in-the-loop**: 4 native speakers, 10 iterations
Multi-lingual experiment: Colors

- new color items extracted from Twitter data:
  - German: 5
  - Italian: 5
  - English: 19
  - Spanish: 22
    - azulgrana
    - rojo vivo
    - “limn” (in place of the color límon)

<table>
<thead>
<tr>
<th></th>
<th>gLD-S</th>
<th>gLD-H</th>
<th>DBSpot.</th>
<th>Babelfy</th>
<th>FRED</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>21</td>
<td>54</td>
<td>13</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>de</td>
<td>18</td>
<td>32</td>
<td>6</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>es</td>
<td>23</td>
<td>43</td>
<td>12</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>it</td>
<td>18</td>
<td>36</td>
<td>8</td>
<td>17</td>
<td>0</td>
</tr>
</tbody>
</table>
Conclusions

WHAT

- knowledge resources are never complete/exhaustive
- construct / improve dictionaries from text corpora

HOW

- iterative and purely statistical algorithm
- no feature extraction required
- comparable behavior for different languages
- organically incorporates human feedback
Multi-lingual Concept Extraction with Linked Data and Human-in-the-Loop

Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, Steve Welch

annalisa.gentile@ibm.com  @AnLiGentile