Language Agnostic Dictionary Extraction

Alfredo Alba[1], Anni Coden[2], Anna Lisa Gentile[1],
Daniel Gruhl, Petar Ristoski[1], and Steve Welch[1]

[1] IBM Research Almaden, CA, US
[2] IBM Watson Research Lab, NY, US
aalba@us.ibm.com, anni@us.ibm.com, annalisa.gentile@ibm.com,
dgruhl@us.ibm.com, petar.ristoski@ibm.com, welchs@us.ibm.com

Abstract. Ontologies are dynamic artifacts that evolve both in structure and content. Keeping them up-to-date is a very expensive and critical operation for any application relying on semantic Web technologies. In this paper we focus on evolving the content of an ontology by extracting relevant instances of ontological concepts from text. The novelty of this work is that we propose a technique which is (i) completely language independent, (ii) combines statistical methods with human-in-the-loop and (iii) exploits Linked Data as bootstrapping source. Experiments on a publicly available parallel medical corpus show comparable performances regardless of the chosen language.

1 Introduction

In this paper we focus on a computer/human partnership to more rapidly evolve the content of an ontology through extraction of new relevant concepts from text. The atomic operation behind this population step is the discovery of all instances that belong to each concept. A plethora of solutions have been proposed to populate ontologies or extract domain dictionaries from both unstructured text [4,6,7,10] and semi-structured content [5,12], but the majority of extraction techniques are language dependent, i.e., they rely on Natural Language Processing (NLP) operations and tools that are language specific, such as parsing, part of speech tagging, etc. We propose *glimpseLD* a novel solution that builds upon our previous work [2] and revolves around three main aspects: (i) it is a statistical method which extracts dictionary items based on context patterns; (ii) it relies on human feedback to automatically tune scores and thresholds for the extraction patterns; (iii) uses Linked Data (when available, even in small quantities) to bootstrap the process. We demonstrate that our approach (i) has similar performances in all languages and that (ii) exploiting Linked Data to bootstrap the method maintains the same comparable performances in all languages, while reducing the number of required human-in-the-loop iterations by at least half.

2 State of the art

There is a vast amount of literature devoted to ontology population from text, with a number of established initiatives to foster research on the topic, such as the Knowledge Base Population task at TAC,[3] the TREC Knowledge Base Acceleration track,[4] and the Open Knowledge Extraction (OKE) Challenge,[5] to name a few. In these initiatives, systems are compared on the basis of recognizing individuals belonging to a few selected ontology classes, spanning from the common Person, Place and Organization,[6] to more specific classes such as Facility, Weapon, Vehicle[7] or Drug,[8] among others. The evaluation focus is usually on the specific sub-tasks involved in the process, such as Entity Recognition, Linking and Typing. Several solutions have been proposed in the literature, spanning from general purpose comprehensive approaches [4] to more domain-specific ones [6,7,10]. The majority of available methods operate (and are assessed) for the English language and although specific initiatives are aimed at encouraging replicable studies in other languages,[9] we argue that truly language-independent methods for this task are not yet widespread and often limited to portability from one language to another [11,9,1]. These methods often exploit linguistic features, which extraction relies on NLP tools - and thus does out-of-the-box portability to different languages is not guaranteed. We propose a human-in-the-loop approach where the human works in partnership with the statistical method to drive the semantic of the task effectively and efficaciously. Moreover we use Linked Data to bootstrap the process. While its usage has been vastly explored for many Information Extraction tasks and specifically for dictionary extraction [3,8,5], the applicability of the models to multiple languages has not been extensively explored.

3 Extracting Dictionaries with *glimpse* and *glimpseLD*

Glimpse is a statistical algorithm for dictionary extraction based on SPOT [2]. The input is a large text corpus whose content is relevant to the domain of the dictionary to be extracted. Besides the corpus, *glimpse* needs one or more examples (*seeds*) of the dictionary items to extract. Starting from these it evaluates the *contexts* (the set of words surrounding an item) in which the seeds occur and identifies “good” contexts to identify further terms or phrases in the corpus, presented to a human to be accepted/rejected (full details of the method can be found in [2]). In this work we synthetically evaluate that *glimpse* is language independent and we prove that using Linked Data to seed the method (*glimpseLD*) can significantly improve the performance, allowing it to extract a higher number of terms in fewer human iterations.

As dataset we use $EMEA$[10] (European Medicines Agency documents), a parallel corpus comprised of PDF documents from the European Medicines Agency,

[3] http://www.nist.gov/tac/2015/KBP

[4] http://trec-kba.org/

[5] https://2016.eswc-conferences.org/eswc-16-open-knowledge-extraction-oke-challenge

[6] http://www.cnts.ua.ac.be/conll2003/ner/

[7] https://www.ldc.upenn.edu/collaborations/past-projects/ace

[8] In Semeval-2013, task 9 https://www.cs.york.ac.uk/semeval-2013/task9/.

[9] Named Entity Recognition and Linking in Italian Tweets: http://www.evalita.it/2016/tasks/neel-it

[10] http://opus.lingfil.uu.se/EMEA.php

related to medical products and their translations into 22 official languages of the European Union. The documents have been sentence aligned within the OPUS project [13]. We select the English, Spanish, Italian and German portion of the dataset and we use it for the task of constructing a dictionary of drugs in the various languages.

From the *EMEA* corpus - and using a standard drug dataset (RxNorm) - we select drugs that appear in all the corpora with the same name. Despite the target terms being the same in all languages, their context is highly language dependent. The selected drugs (363 in total) are consider as Gold Standard and we start with one seed only in every language (specifically we used the drug *irbesartan*) to automatically build a drug dictionary in every target language. The behavior of *glimpse* is homogeneous in every language, with similar discovery growth at each iteration. With 20 iterations *glimpse* discovered more than 300 drugs in each language (out of the 363 of the GS, with average accuracy [11] >85%). The average Pearson correlation amongst the results in all languages is > 0.99. The *discovery growth* - the ratio of new correct terms added - is a useful indication of performance in a real scenario, where no gold standard is available, but correctness of extraction is assured by human-in-the-loop.

We repeat the experiment with *glimpseLD*. We build a truly multi-language GS crawling instances of drugs from Linked Data, making sure to cover the same drugs in all languages. Particularly, we use two of the biggest cross-domain LOD datasets, DBpedia[12] and Wikidata.[13] We select all the entities of type *dbo:Drug*[14] from DBpedia and all the entities of type *wikidata:Q11173* from Wikidata. For all of the selected entities, we retrieve the corresponding labels in English, German, Spanish and Italian and consider this our gold standard dictionary. We then select 20% as seeds and measure the performance of recreating the remaining 80% by using *glimpseLD*. We perform 5-fold cross validation without repetition and randomly select the 20% of seeds at each iteration (making sure that the seeds represent the same drugs for all 4 languages), to test if the choice of initial seeds impacts the results. Fig. 1b and 1c show that the discovery growth is comparable for all languages, with correlation always above 0.98.

4 Conclusions and future work

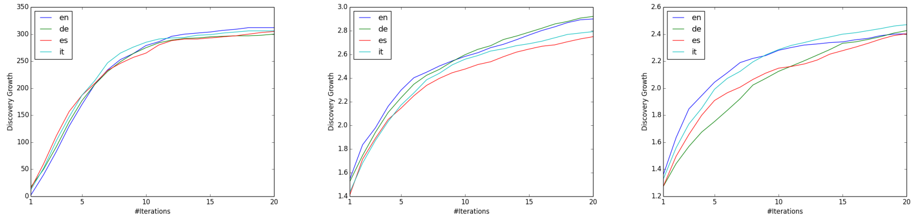
This paper proposes a language-independent solution to discover new instances for populating ontology concepts. Our algorithm is iterative and purely statistical, hence does not require any feature extraction which can be difficult and expensive in different languages and texts. It leverages Linked Data to seed the process, and integrates human feedback to improve the accuracy and the control concept drift at every iteration cycle. We show extremely similar discovery growth extracting drug names on four languages over parallel corpora of medical text.

[11] Note that as irrelevant terms are manually rejected in a human-in-the-loop approach it does not make sense to calculate Precision, as retained terms are all correct by design.

[12] http://.dbpedia.org

[13] https://www.wikidata.org/

[14] *dbo*: http://dbpedia.org/ontology/, *wikidata*: http://www.wikidata.org/entity/



a) *r = 0.998.* **b)** *r = 0.997* **c)** *r= 0.985*

Fig. 1: Comparison of *discovery growth* for *glimpse* and *glimpseLD* across different languages on the EMEA dataset, using one manual seed (Fig 1a), seeds from DBpedia (Fig 1b) and Wikidata (Fig 1c). Pearson correlation (r) amongst results from different languages is reported.

References

1. A. Ben Abacha, M. F. M. Chowdhury, A. Karanasiou, Y. Mrabet, A. Lavelli, and P. Zweigenbaum. Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification. *Journal of Biomedical Informatics*, 58:122–132, 2015.
2. A. Coden, D. Gruhl, N. Lewis, M. Tanenblatt, and J. Terdiman. SPOT the drug! An unsupervised pattern matching method to extract drug names from very large clinical corpora. *HISB 2012*, pages 33–39, 2012.
3. J. Dolby, A. Fokoue, A. Kalyanpur, E. Schonberg, and K. Srinivas. Extracting enterprise vocabularies using linked open data. *ISWC 2009*, pages 779–794, 2009.
4. A. Gangemi, V. Presutti, D. Reforgiato Recupero, A. G. Nuzzolese, F. Draicchio, and M. Mongiovì. Semantic web machine reading with fred. *Semantic Web*, (Preprint):1–21, 2016.
5. A. L. Gentile, Z. Zhang, I. Augenstein, and F. Ciravegna. Unsupervised wrapper induction using linked data. In *K-CAP'13*, pages 41–48. ACM, 2013.
6. K. Lee, A. Qadir, S. A. Hasan, V. Datla, A. Prakash, J. Liu, and O. Farri. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *WWW'17*, pages 705–714, 2017.
7. S. Liu, B. Tang, Q. Chen, and X. Wang. Effects of semantic features on machine learning-based drug name recognition systems: Word embeddings vs. Manually constructed dictionaries. *Information (Switzerland)*, 6(4):848–865, 2015.
8. P. Mitzias, M. Riga, E. Kontopoulos, T. G. Stavropoulos, S. Andreadis, G. Meditskos, and I. Kompatsiaris. User-driven ontology population from linked data sources. In *KESW 2016*, pages 31–41. Springer, 2016.
9. A. Pappu, R. Blanco, Y. Mehdad, A. Stent, and K. Thadani. Lightweight multilingual entity extraction and linking. In *WSDM '17*, pages 365–374. ACM, 2017.
10. N. Pröllochs, S. Feuerriegel, and D. Neumann. Generating Domain-Specific Dictionaries using Bayesian Learning. *Ecis*, (2015):0–14, 2015.
11. M. Sahlgren and J. Karlgren. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Nat. Lang. Eng.*, 11(3):327–341, Sept. 2005.
12. H.-J. Song, S.-B. Park, and S.-Y. Park. An automatic ontology population with a machine learning technique from semi-structured documents. In *ICIA'09*, pages 534–539. IEEE, 2009.
13. J. Tiedemann. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. *RANLP*, 2009.