TALN-UPF in Computational Lexicography

Luis Espinosa-Anke[1]

[1]TALN-DTIC, Universitat Pompeu Fabra, Carrer Tànger, 122-134, Barcelona (Spain)
luis.espinosa@upf.edu

1. Introduction

In this paper, we provide a survey of prominent work carried out in the intersection between Natural Language Processing (NLP) and Lexicography by the TALN Natural Language Processing group, at Pompeu Fabra University in Barcelona (Spain). First, we focus on *Lexicography for NLP*, i.e. taking advantage of lexicographic resources in tasks such as lexical semantics and lexical taxonomy learning. Second, we cover the area of *NLP for Lexicography*, where we tackle current and past projects where NLP techniques are applied to ease up the lexicographic process. This includes, generally, the automatic acquisition of lexicographic items such as textual definitions, hypernyms or collocations from naturally occurring text.

2. Lexicography for NLP

The contribution of lexicographic sources in NLP has not received much attention recently. This may be justified due to the increasing number of tasks where neural approaches have shown to outperform traditional methods, and hence the focus is shifting from structured feature engineering to the best parameter tuning in a neural network architecture. However, the structured, curated and high-quality information that can be derived from lexicographic resources, we argue, can contribute dramatically to the performance of algorithms designed to solve an NLP task. In this paper we focus on two tasks where *lexicographic* resources were a core component, namely discovering hypernymic relations (which are useful for reasoning or semantic search), and automatic taxonomy learning.

Hypernym Detection

The Aristotelian *genus-et-differentia* model of a definition [Storrer and Wellinghoff, 2006] defines a class-superclass or hypernymic relation between the *definiendum* (the concept that is being defined), and the *genus* (the superclass, contained in the cluster of words that define the *definiendum*, called *definiens*). Extracting hypernymic relations can be useful not only for the web, Semantic Search, Question Answering or Reasoning, but also as a core component of a semi-automatic glossary or dictionary construction process. Given these potential fields of application, there has been substantial interest in capturing hypernymic relations from corpora. Textual definitions following the genus-et-differentia structure are rich in encoding (term, hypernym) relations. For example, in the following case: "A mosque/`Term` is a building/`Hypernym` where muslims go to pray". This is not a trivial task, since even within their stylistic constraints, textual definitions can be expressed in a plethora of ways, and hence rule-based methods based on pattern matching are unsuitable.

We proposed a machine learning-based algorithm in [Espinosa-Anke et al., 2015a], evaluated on a standard dataset for definition-level hypernym detection, the WCL dataset [Navigli et al., 2010], a corpus of manually annotated Wikipedia definitions along with the term's hypernym. It outperformed the state of the art at the time [Navigli and Velardi, 2010,Boella et al., 2014], and proved the effectiveness of combining linguistically motivated features with machine learning approaches.

Taxonomy Learning

Another relevant project is *ExTaSem!* [Espinosa-Anke et al., 2016], where we leverage the vast definitional information contained in BabelNet [Navigli and Ponzetto, 2010] together with its integrated two-layered taxonomy [Flati et al., 2014] to develop an algorithm for unsupervised domain-specific taxonomy learning. Our algorithm takes as input a list of terms in a given domain (e.g. Food, Equipment or Chemical), and returns a fully disambiguated lexical taxonomy (i.e. a directed acyclic graph where each node represents a concept and each edge, an is-a relation between them). It is trained on the WCL dataset, and with a set of features described and evaluated in [Espinosa-Anke et al., 2015b], we trained a Conditional Random Fields [Lafferty et al., 2001] sequential classifier which extracted thousands of (term, hypernym) pairs. After several steps which we do not detail due to space constraints, *ExTaSem!* produces a taxonomy with an extended vocabulary of several orders of magnitude larger than the original terminology, and disambiguates most of its nodes against BabelNet with high accuracy. Somewhat related is our follow-up project, *TaxoEmbed* (recently accepted as long paper in EMNLP 2016), which also takes advantage of definitional information in BabelNet, but this time indirectly, as definitions for BabelNet synsets provide key information for constructing *synset-level* vectors which we use in our experiments. We use vector space information to (1) cluster BabelNet synsets by *domain* (e.g. Music, Tranport and Travel or Sports); and (2) train an algorithm specific to these domains so that it is capable of reliably constructing a disambiguated lexical taxonomy in these domains. Our evaluation, which is carried out in Wikidata, shows that only with distributional and structural (i.e. definitional) information it is possible to rival the quality of hand-crafted Wikidata hypernymic relations.

3. NLP for Lexicography

Contributions of NLP to Lexicography are usually better defined that vice versa, mostly because of the direct application of certain systems to easing up certain lexicographic tasks like finding suitable definitions or examples in corpora. In our group, we have focused in the last years mostly in two aspects: First, the automatic extraction of textual definitions from corpora, foreseeing an application in making it easier for lexicographers to build domain glossaries. Second, the automatic extraction, classification, correction and visualization of collocations.

Definition Extraction

In [Espinosa-Anke and Saggion, 2014], we proposed a supervised machine learning approach to discover definitions from text. We set up our experiments so that we assessed the extent to which our algorithm was able to distinguish between a canonical textual definition (embodied as the first

sentence of randomly sampled Wikipedia articles), and distractors. These distractors are what the authors of the dataset [Navigli et al., 2010] defined as "syntactically plausible false definitions". In our experiments, we showed that we were able to outperform the state of the art in this dataset [Navigli and Velardi, 2010,Boella et al., 2014] by combining linguistically motivated features stemming from the linguistic formalism of dependency grammar, and training with Support Vector Machines. Furthermore, we proposed an additional algorithm which incorporated features derived from *sense-level* word embeddings [Iacobacci et al., 2015]. We modelled each candidate definition as a graph, where each node represents a sense, and each edge is weighted according to the pair of nodes' cosine similarity according to their corresponding vectors [Espinosa-Anke et al., 2015]. Finally, we have also explored *semi-supervised* definition extraction systems, which are capable to adapting gradually to a target corpus. Our experiments are evaluated in [Espinosa-Anke et al., 2015b], and released as open source software [Espinosa-Anke et al., 2016][1].

Collocation Processing

Let us cover past and current projects involving collocation discovery. An ongoing project in which we are currently working is the extension of WordNet [Miller, 1995] with collocational information, as described in the Meaning Text Theory [Mel'čuk, 1996]. Succintly, our goal is to, given the pair of synsets `desire.n.01` and `ardent.a.01`, to encode a novel relation $\xrightarrow[x]{col:intense}$ between them, where 'intense' is the *semantic category* denoting *intensification*, and x is the confidence score assigned by our algorithm. The derived resource, furthermore, can be used to *retrofit* any word embeddings model, strengthening the relation existing between collocation items. This can be useful for automatic collocation discovery or for offering alternative *collocates* (in our previous example, alternatives to *heavy*, if applicable) for a given *base* (in our example, *rain*). This work extends previous and successful approaches in collocation acquisition from text corpora, both unsupervised [Rodríguez-Fernández et al., 2016a] and supervised [Rodríguez-Fernández et al., 2016b].

Moreover, a recent paper published in the International Journal of Lexicography focuses on the automatic classification of collocations using machine learning techniques [Wanner et al., 2016]. In their work, Wanner et al. propose a combination of lexical, morphological and morphosyntactic features, where each collocation is modelled according to relevant ngrams it appears with, typical collocates, and typical syntactic relations between base and collocate according their syntactic dependencies. This is an extension of previous work carried out in these lines, e.g. [Wanner et al., 2006].

In addition to automating the acquisition of collocation resources, another relevant line of research at TALN is automatic collocation error detection. As of today, we are exploring neural networks approaches which cast the task as a sequential labelling problem, tagging candidate text spans as correct or incorrect collocations. This, however, is still in an early stage, and thus we do not fully elaborate on this. Still, let us refer to the work by [Ferraro et al., 2014], where collocation errors are tackled by considering the *association strength* (co-occurrence) of a base and its collocate, together with other criteria such as character bigrams (to study the overlap between correct and incorrect collocations). The authors also applied a *lexical context metric*, which takes into account the context in which a miscollocation occurs, much in the current line of distributional semantics, where words with similar meanings are expected to share similar contexts.

[1] https://bitbucket.org/luisespinosa/defext

The last project involving collocations we cover consists on the improvement of electronic collocation resources by enhancing them with visual analytics techniques [Carlini et al., 2015]. The authors show different ways to explore collocational resources. In their platform, for instance, it is possible to investigate the collocation space of a base (i.e. *rain* in *heavy rain*), or the collocation space of bases sharing collocates.

4. Conclusion

We have provided a brief overview of some of the most prominent works carried out at the TALN research group in the intersection between lexicography and NLP. Today, with lexicographic sources being in many cases collective efforts, and in many cases compliant with Linguistic Linked Data standards, there is an increasing interest in taking advantage of the knowledge they encode for downstream applications in NLP and Artificial Intelligence. Conversely, these fields can contribute dramatically to the lexicographic process thanks to the availability of systems, for instance, capable of extracting and classifying definitions or collocations from corpora.

References

Boella et al., 2014. Boella, G., Di Caro, L., Ruggeri, A., and Robaldo, L. (2014). Learning from syntax generalizations for automatic semantic annotation. *Journal of Intelligent Information Systems*, pages 1–16.

Carlini et al., 2015. Carlini, R., Codina-Filbà, J., and Wanner, L. (2015). Improving the use of electronic collocation resources by visual analytics techniques. In *Proceedings of the eLex 2015 Conference*.

Espinosa-Anke et al., 2016. Espinosa-Anke, L., Carlini, R., Ronzano, F., and Saggion, H. (2016). Defext: A semi supervised definition extraction tool. In *Proceedings of GLOBALEX*, Portoroz, Slovenia.

Espinosa-Anke and Saggion, 2014. Espinosa-Anke, L. and Saggion, H. (2014). Applying dependency relations to definition extraction. In *Natural Language Processing and Information Systems*, pages 63–74. Springer.

Espinosa-Anke et al., 2015. Espinosa-Anke, L., Saggion, H., and Delli Bovi, C. (2015). Definition extraction using sense-based embeddings. In *Proceedings of the 2015 International Workshop on Embeddings and Semantics*, pages 10–15, Alicante, Spain.

Espinosa-Anke et al., 2015a. Espinosa-Anke, L., Saggion, H., and Ronzano, F. (2015a). Hypernym extraction: Combining machine learning and dependency grammar. In *CICLING 2015*, page To appear, Cairo, Egypt. Springer-Verlag.

Espinosa-Anke et al., 2015b. Espinosa-Anke, L., Saggion, H., and Ronzano, F. (2015b). Weakly supervised definition extraction. In *Proceedings of RANLP 2015*, pages 176–185.

Espinosa-Anke et al., 2016. Espinosa-Anke, L., Saggion, H., Ronzano, F., and Navigli, R. (2016). Extasem! extending, taxonomizing and semantifying domain terminologies. AAAI.

Ferraro et al., 2014. Ferraro, G., Nazar, R., Ramos, M. A., and Wanner, L. (2014). Towards advanced collocation error correction in spanish learner corpora. *Language Resources and Evaluation*, 48(1):45–64.

Flati et al., 2014. Flati, T., Vannella, D., Pasini, T., and Navigli, R. (2014). Two is bigger (and better) than one: the wikipedia bitaxonomy project. In *ACL*.

Iacobacci et al., 2015. Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105, Beijing, China.

Lafferty et al., 2001. Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mel'čuk, 1996. Mel'čuk, I. (1996). Lexical functions: A tool for the description of lexical relations in the lexicon. In Wanner, L., editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. Benjamins Academic Publishers, Amsterdam.

Miller, 1995. Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Navigli and Ponzetto, 2010. Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *ACL*, pages 216–225.

Navigli and Velardi, 2010. Navigli, R. and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *ACL*, pages 1318–1327.

Navigli et al., 2010. Navigli, R., Velardi, P., and Ruiz-Martínez, J. M. (2010). An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of LREC'10*, Valletta, Malta.

Rodríguez-Fernández et al., 2016a. Rodríguez-Fernández, S., Carlini, R., Espinosa-Anke, L., and Wanner, L. (2016a). Example-based acquisition of fine-grained collocation resources. In *Proceedings of LREC*, Portoroz, Slovenia.

Rodríguez-Fernández et al., 2016b. Rodríguez-Fernández, S., Espinosa-Anke, L., Carlini, R., and Wanner, L. (2016b). Semantics-driven recognition of collocations using word embeddings. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, Berlin, Germany.

Storrer and Wellinghoff, 2006. Storrer, A. and Wellinghoff, S. (2006). Automated detection and annotation of term definitions in German text corpora. In *Conference on Language Resources and Evaluation (LREC)*.

Wanner et al., 2006. Wanner, L., Bohnet, B., and Giereth, M. (2006). Making sense of collocations. *Computer Speech & Language*, 20(4):609–624.

Wanner et al., 2016. Wanner, L., Ferraro, G., and Moreno, P. (2016). Towards distributional semantics-based classification of collocations for collocation dictionaries. *International Journal of Lexicography*, page ecw002.