The interplay between lexical resources and Natural Language Processing

Abstract

Incorporating linguistic, world and common sense knowledge into AI/NLP systems is currently an important research area, with several open problems and challenges. At the same time, processing and storing this knowledge in lexical resources is not a straightforward task. We propose to address these complementary goals from two methodological perspectives: the use of NLP methods to help the process of constructing and enriching lexical resources and the use of lexical resources for improving NLP applications. This tutorial may be useful for two main types of audience: those working on language resources who are interested in becoming acquainted with automatic NLP techniques, with the end goal of speeding and/or easing up the process of resource curation; and on the other hand, researchers in NLP who would like to benefit from the knowledge of lexical resources to improve their systems and models.

1. Description

The manual construction of lexical resources is a prohibitively time-consuming process, and even in the most restricted knowledge domains and less-resourced languages, the use of language technologies to ease up this process is becoming a standard practice. NLP techniques can be effectively leveraged to reduce creation and maintenance efforts. In this tutorial we will present open problems and research challenges in these topics concerning the interplay between lexical resources and NLP. Additionally, we will summarize existing attempts in this direction, such as modeling linguistic phenomena like terminology, definitions and glosses, examples and relations, phraseological units, or clustering techniques for senses and topics, as well as the integration of resources of different nature. The following topics are going to be covered in detail:

- **Terminology extraction.** Measures for terminology extraction, the simple conventional tf-idf (Sparck Jones, 1972), lexical specificity (Lafon, 1980), and more recent approaches exploiting linguistic knowledge (Hulth 2003; Vivaldi and Rodríguez, 2010).

- **Definition extraction.** Techniques for extracting definitional text snippets from corpora (Navigli and Velardi, 2010; Boella and DiCaro, 2013; Espinosa-Anke et al. 2015; Li et al. 2016).

Proceedings of NAACL-HLT 2018: Tutorial Abstracts, pages 17–23
New Orleans, Louisiana, June 1, 2018.

- **Automatic extraction of examples.** Description of example extraction techniques and designs on this direction, e.g., the GDEX criteria and their implementation (Kilgariff et al., 2008).

- **Information extraction.** Recent approaches for extracting semantic relations from text: NELL (Carlson et al., 2010), ReVerb (Fader et al. 2011), PATTY (Nakashole et al., 2011), KB-Unify (Delli Bovi et al., 2015).

- **Hypernym discovery and taxonomy learning.** Insights from recent SemEval tasks (Bordea et al. 2015, 2016) and related efforts on the automatic extraction of hypernymy relations from text corpora (Velardi et al. 2013; Alfarone and Davis 2015; Flati et al. 2016; Shwartz et al. 2016; Espinosa-Anke et al. 2016a; Gupta et al. 2016).

- **Topic clustering techniques**. Relevant techniques for filtering general domain resources via topic grouping (Roget's, 1911; Navigli and Velardi, 2004, Camacho-Collados and Navigli, 2017).

- **Alignment of lexical resources:** Alignment of heterogeneous lexical resources contributing to the creation of large resources containing different sources of knowledge. We will present approaches for the construction of such resources, such as Yago (Suchanek et al. 2007), UBY (Gurevych et al. 2012), BabelNet (Navigli and Ponzetto, 2012) or ConceptNet (Speer et al. 2017), as well as other works attempting to improve the automatic procedures to align lexical resources (Matuschek and Gurevych, 2013; Pilehvar and Navigli, 2014).

- **Ontology enrichment.** Enriching lexical ontologies with novel synsets or with additional relations (Jurgens and Pilehvar, 2015; 2016; Espinosa-Anke et al., 2016b).

In addition to these automatic efforts for easing the task of constructing and enriching lexical resources, we will present NLP tasks in which lexical resources have shown an important contribution. Effectively leveraging linguistically expressible cues with their associated knowledge remains a difficult task. Knowledge may be extracted from (roughly) three types of resources (Hovy et al., 2013): unstructured, e.g. text corpora; semistructured, such as encyclopedic collaborative repositories like Wikipedia and Wiktionary, or structured, which include lexicographic resources like WordNet or DBpedia.

We will explain some of the current challenges in Word Sense Disambiguation and Entity Linking, as key tasks in natural language understanding which also enable a direct integration of knowledge from lexical resources. We will explain some knowledge-based and supervised methods for these tasks which play a decisive role in connecting lexical resources and text data (Zhong and Ng, 2010; Agirre et al. 2014; Moro et al.. 2014; Ling et al. 2015; Raganato et al. 2017). Moreover, we will present the field of knowledge-based representations, in particular word sense embeddings (Chen et al. 2014; Rothe and Schuetze, 2015; Camacho-Collados et al. 2016; Pilehvar and Collier, 2016; Mancini et al. 2017), as flexible techniques which act as a bridge between lexical resources and

applications. Finally, we will briefly present some recent work on the integration of this encoded knowledge from lexical resources into neural architectures for improving downstream NLP applications (Flekova and Gurevych, 2016; Pilehvar et al. 2017).

2. Outline

➢ Introduction and Motivation (15 mins)

Adding explicit knowledge into AI/NLP systems is currently an important challenge due to the gains that can be obtained in many downstream applications. At the same time, these resources can be further enriched and better exploited by making use of NLP techniques. In this context, the main motivation of this tutorial is to show how Natural Language Processing and Lexical Resources have interacted so far, and a view towards potential scenarios in the near future.

The tutorial is then divided in two main blocks. First, we delve into *NLP for Creation and Enrichment of Lexical Resources*, where we address a range of NLP problems aimed specifically at improving repositories of linguistically expressible knowledge. Second, we cover different use cases in which *Lexical Resources for NLP* have been leveraged successfully. The last part of the tutorial focuses on lessons learned from work in which we tried to reconcile both worlds, as well as our own view towards what the future holds for knowledge-based approaches to NLP.

➢ NLP for Lexical Resources (70 mins)

The application of language technologies to the automatic construction and extension of lexical resources has proven successful in that it has provided various tools for optimizing this often prohibitively costly and expensive process.. NLP techniques provide end-to-end technologies that can tackle all challenges in the language resource creation and maintenance pipeline.. In this tutorial we will summarize existing efforts in this direction, including the extraction from text of linguistic phenomena like terminology, definitions and glosses, examples and relations, as well as clustering techniques for senses and topics. We will additionally summarize recent work on the automatic integration of knowledge from heterogeneous resources such as BabelNet, ConceptNet, Uby or Yago.

[Coffee break] (20 mins)

➢ Lexical Resources for NLP (60 mins)

In this section we will present some of the applications on which lexical resources play an important role. In particular, we will explain some of the problems and challenges in Word Sense Disambiguation and Entity Linking, as key tasks in natural language understanding. Moreover, we will present the field of knowledge-based representations, in particular sense vectors and embeddings, as flexible techniques connecting lexical resources and downstream applications. We will additionally present some recent works on the integration of knowledge-based embeddings into neural architectures for improving downstream NLP applications.

➢ Open problems and challenges (15 mins)

In this last section we will introduce some of the open problems and challenges for automatizing the resource creation and enrichment process as well as for the integration of knowledge from lexical resources into NLP applications.

3. Instructors

Jose Camacho Collados
(camachocolladosj@cardiff.ac.uk; http://www.josecamachocollados.com) is a Research Associate at Cardiff University. Previously he was a Google Doctoral Fellow and completed his PhD at Sapienza University of Rome. His research focuses on Natural Language Processing and, more specifically, on the area of lexical and distributional semantics. Jose has experience in utilizing lexical resources for NLP applications, while enriching and improving these resources by extracting and processing knowledge from textual data. On this area he has co-organized the SemEval 2018 shared task on Hypernym Discovery. Previously, he co-organized a workshop on “Sense, Concept and Entity Representations and their Applications” at EACL 2017 and a tutorial on the same topic at ACL 2016. His background education includes an Erasmus Mundus Master in Natural Language Processing and Human Language Technology and a 5-year BSc degree in Mathematics.

Luis Espinosa Anke
(espinosa-ankel@cardiff.ac.uk, www.luisespinosa.net) received his BA in English Philology in 2006 (Univ. of Alicante, Spain), and his PhD in Natural Language Processing in 2017 (Univ. Pompeu Fabra, Spain). He holds two MAs, one in English-Spanish Translation (Univ. of Alicante), and an Erasmus Mundus MA in Natural Language Processing (NLP) (Univ. of Wolverhampton and Univ. Autònoma de Barcelona). His research interests lie in the intersection between structured representations of knowledge and NLP, specifically computational lexicography and distributional semantics. He has co-organized the SemEval 2018 shared tasks on Hypernym Discovery and Multilingual Emoji Prediction. Previously, he co-organized the Spanish NLP conference (2014) and the Focused NER task (Open Knowledge Extraction challenge) at ESWC 2017.

Mohammad Taher Pilehvar
(mp792@cam.ac.uk, http://people.ds.cam.ac.uk/mp792/) is a research associate at the University of Cambridge. Taher's research lies in lexical semantics, mainly focusing on semantic representation and similarity. In the past, he has co-instructed three tutorials on these topics (EMNLP 2015, ACL 2016, and EACL 2017) and co-organised three SemEval tasks. He has also co-authored several conference (including two ACL best paper nominations, at 2013 and 2017) and journal papers, including different semantic representation techniques based on heterogeneous lexical resources.

References

Agirre, E., de Lacalle, O.L. and Soroa, A., 2014. Random walks for knowledge-based word sense disambiguation. Computational Linguistics, 40(1), pp.57-84.

Boella, G. and Di Caro, L., 2013, August. Extracting Definitions and Hypernym Relations relying on Syntactic Dependencies and Support Vector Machines. In ACL (2) (pp. 532-537).

Camacho-Collados, J., Pilehvar, M. T., & Navigli, R., 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. Artificial Intelligence, 240, 36-64.

Camacho-Collados, J. and Navigli, R., 2017. BabelDomains: Large-Scale Domain Labeling of Lexical Resources. EACL 2017, p.223.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E.R. and Mitchell, T.M., 2010, July. Toward an Architecture for Never-Ending Language Learning. In AAAI (Vol. 5, p. 3).

Chen, X., Liu, Z., & Sun, M., 2014. A Unified Model for Word Sense Representation and Disambiguation. In Proceedings of EMNLP (pp. 1025-1035).

Delli Bovi, C., Espinosa-Anke, L. and Navigli, R., 2015. Knowledge base unification via sense embeddings and disambiguation. In The 2015 Conference on Empirical Methods in Natural Language; Lisbon, Portugal. pp. 726-36.

Espinosa-Anke, L., Saggion, H. and Ronzano, F., 2015. Weakly supervised definition extraction. In Proceedings of the International Conference Recent Advances in Natural Language Processing (pp. 176-185).

Espinosa-Anke, L., Camacho-Collados, J., Bovi, C. D., & Saggion, H. (2016). Supervised distributional hypernym discovery via domain adaptation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 424-435).

Espinosa-Anke, L., Camacho-Collados, J., Rodríguez-Fernández, S., Saggion, H., & Wanner, L. (2016). Extending WordNet with Fine-Grained Collocational Information via Supervised

Distributional Learning. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 3422-3432).

Fader, A., Soderland, S. and Etzioni, O., 2011. Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535-1545). Association for Computational Linguistics.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., & Wirth, C., 2012. Uby: A large-scale unified lexical-semantic resource based on LMF. In Proceedings of EACL (pp. 580-590).

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. and Rychlý, P., 2008, July. GDEX: Automatically finding good dictionary examples in a corpus. In Proc. Euralex.

Flekova, L., & Gurevych, I., 2016. Supersense Embeddings: A Unified Model for Supersense Interpretation, Prediction, and Utilization. In Proceedings of ACL.

Hovy, E., Navigli, R. and Ponzetto, S.P., 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence, 194, pp.2-27.

Hulth, A., 2003. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 216-223). Association for Computational Linguistics.

Jurgens, D., & Pilehvar, M. T. (2015). Reserating the awesometastic: An automatic extension of the WordNet taxonomy for novel terms. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1459-1465).

Jurgens, D., & Pilehvar, M. T. (2016). Semeval-2016 task 14: Semantic taxonomy enrichment. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 1092-1102).

Lafon, P., 1980. Sur la variabilité de la fréquence des formes dans un corpus. Mots, 1(1), pp.127-165.

Mancini, M., Camacho-Collados, J., Iacobacci, I., & Navigli, R., 2016. Embedding Words and Senses Together via Joint Knowledge-Enhanced Training. In Proceedings of CoNLL 2017.

Matuschek, M. & Gurevych, I., 2013. Dijkstra-wsa: A graph-based approach to word sense alignment. In Transactions of the Association for Computational Linguistics, 1, pp.151-164

Navigli, R. and Velardi, P., 2010. Learning word-class lattices for definition and hypernym extraction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 1318-1327). Association for Computational Linguistics.

Navigli, R., & Ponzetto, S. P., 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193, 217-250.

Pilehvar, M. T., & Collier, N., 2016. De-conflated semantic representations. In Proceedings of EMNLP.

Pilehvar, M. T., Camacho-Collados, J., Navigli, R., & Collier, N., 2017. Towards a Seamless Integration of Word Senses into Downstream NLP Applications. Proceedings of ACL.

Pilehvar, M.T. & Navigli, R., 2014. A robust approach to aligning heterogeneous lexical resources. In Proceedings of ACL.

Rothe, S., & Schütze, H., 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In Proceedings of ACL.

Spark-Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(5), pp.111-121.

Speer, R., Chin, J. and Havasi, C., 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In AAAI (pp. 4444-4451).

Suchanek, F. M., Kasneci, G., & Weikum, G., 2007. Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (pp. 697-706). ACM.

Vivaldi, J. and Rodríguez, H., 2010. Finding Domain Terms using Wikipedia. In LREC.

Zhong, Z. and Ng, H.T., 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In Proceedings of the ACL 2010 System Demonstrations (pp. 78-83).