Combining Syntactic and Sequential Patterns for Unsupervised Semantic Relation Extraction

Nadège Lechevrel[1], Kata Gábor[2], Isabelle Tellier[3], Thierry Charnois[2], Haïfa Zargayouna[2], Davide Buscaldi[2],

[1] Université Paris-Ouest Nanterre La Défense
[2] LIPN, CNRS (UMR 7030), Université Paris 13
[3] LaTTiCe, CNRS (UMR 8094), ENS Paris, Université Sorbonne Nouvelle - Paris 3
PSL Research University, Université Sorbonne Paris Cité

This work investigates the impact of syntactic features in a completely unsupervised semantic relation extraction experiment. Automated relation extraction deals with identifying semantic relation instances in a text and classifying them according to the type of relation. This task is essential in information and knowledge extraction and in knowledge base population. Supervised relation extraction systems rely on annotated examples [13, 23–25, 29] and extract different kinds of features from the training data, and eventually from external knowledge sources. The types of extracted relations are necessarily limited to a pre-defined list. In *Open Information Extraction (OpenIE)* [1, 4] relation types are inferred directly from the data: concept pairs representing the same relation are grouped together and relation labels can be generated from context segments or through labeling by domain experts [1, 5, 6]. A commonly used method [21, 22] is to represent entity couples by a *pair-pattern matrix*, and cluster relation instances according to the similarity of their distribution over patterns. Pattern-based approaches [2,11, 21, 23, 26] typically use lexical context patterns, assuming that the semantic relation between two entities is explicitly mentioned in the text. Patterns can be defined manually [11], obtained by Latent Relational Analysis [21], or from a corpus by sequential pattern mining [2, 9, 20]. Previous works, especially in the biomedical domain, have shown that not only lexical patterns, but also syntactic dependency trees can be beneficial in supervised and semi-supervised relation extraction [3, 7, 17–19]. Early experiments on combining lexical patterns with different types of distributional information in unsupervised relation clustering did not bring significant improvement [12]. The underlying difficulty is that while supervised classifiers can learn to weight attributes from different sources, it is not trivial to combine different types of features in a single clustering feature space.

In our experiments, we propose to combine syntactic features with sequential lexical patterns for unsupervised clustering of semantic relation instances in the context of (NLP-related) scientific texts. We replicate the experiments of [9] and augment them with dependency-based syntactic features. We adopt a pair-pattern matrix for clustering relation instances. The task can be described as follows: if a_1, a_2, b_1, b_2 are pre-annotated domain concepts extracted from a corpus, we would like to classify concept pairs $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in homogeneous groups according to their semantic relation. We need an efficient

In: P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens, Y. Toussaint (Eds.): Proceedings of DMNLP, Workshop at ECML/PKDD, Skopje, Macedonia, 2017.

representation of (a_1, a_2) and (b_1, b_2) in a vector space which allows to calculate relational similarity $sim(a, b)$ and cluster concept pairs. Concept couples were extracted from ACL-RelAcs corpus [8] based on frequency of co-occurrence in the same sentence. In [9,10] a set of entities were manually categorized in 20 semantic relations; we use this dataset in the clustering and for evaluation. Both sequential patterns and syntactic features are extracted automatically from the same corpus. The input pairs were first represented using the best performing pattern representation in [9], i.e. sequential patterns. A sequence, in this context, is a list of literals (items) where an item is a word in the corpus. Only *closed* sequential patterns were considered, i.e. patterns which are not sub-sequences of another sequence with the same support. This feature space is then augmented using syntactic patterns. A dependency parsing approach was adopted: the structure of sentences is described in terms of binary relations between words, where each word depends on another one. A dependency structure consists in a set of triplets where two lexical items are linked by a typed arc indicating the nature of their grammatical relationship. The dependency structures are thus labeled directed graphs consisting of a set of vertices V (the set of words/punctuation in a given sentence) and a set of pairs of vertices A (the arcs and their types which correspond to the grammatical relationships between the elements in V): G = (V,A). Each dependency structure is a set of triplets where two lexical items are linked by a typed arc indicating the nature of their grammatical relationship. In our experiments, we used the Stanford dependency scheme [16], a semantically-oriented dependency representation. The parser is Stanford Parser version 3.8.0, trained on the Penn Treebank [14]. The basic typed dependencies representation (see [15] for a description of the labels) was chosen. Following [3], who have shown that the shortest path between two entities in the dependency tree can be used to improve relation extraction, the shortest paths between two eligible entities is extracted along with the grammatical information contained in the types. The dependency label sequences of the shortest path were transformed into attributes. The experiments aimed at comparing clustering results based on sequential patterns alone, syntactic information alone, and a mixed representation with both types of information. Three feature spaces were thus constructed: 1) the sequential pattern attributes of [9], 2) the dependency path attributes and 3) a combined feature space using both types of attributes. Clustering was done using a hierarchical agglomerative clustering with a bisective initialization [27] implemented in cluto [28]. The initialization by repeated bisections yields a number of centroids that serve to augment the original feature space; the values of these new dimensions are given by the distance of each object (here: concept pairs) from the centroids. The clusters were evaluated against the manually categorized sample of [9]. Results show that the combination of both information is beneficial: the feature space 3) provides the best clusters.

Acknowledgments This work is part of the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL).

References

1. M. Banko, J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI*, 2007.
2. N. Béchet, P. Cellier, T. Charnois, and B. Crémilleux. Discovering linguistic patterns using sequence mining. In *CICLing '12*, 2012.
3. R. C. Bunescu and R. J. Mooney. A shortest path dependency kernel for relation extraction. In *HLT-EMNLP'05*, 2005.
4. L. Del Corro and R. Gemulla. Clausie: Clause-based open information extraction. In *International Conference on World Wide Web*, WWW '13, 2013.
5. A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP '11*, 2011.
6. O. Ferret. *Language Production, Cognition, and the Lexicon*, chapter Typing Relations in Distributional Thesauri, pages 113–134. Springer International Publishing, 2015.
7. K. Fundel, R. Küffner, and R. Zimmer. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365, 2007.
8. K. Gábor, H. Zargayouna, D. Buscaldi, I. Tellier, and T. Charnois. Semantic annotation of the acl anthology corpus for the automatic analysis of scientific literature. In *LREC '16*, 2016.
9. K. Gábor, H. Zargayouna, D. Buscaldi, I. Tellier, and T. Charnois. Unsupervised relation extraction in specialized corpora using sequence mining. In *Advances in Intelligent Data Analysis XV (IDA 2016), LNCS 9897*, 2016.
10. K. Gábor, H. Zargayouna, I. Tellier, D. Buscaldi, and T. Charnois. A typology of semantic relations dedicated to scientific literature analysis. In *SAVE-SD Workshop at the 25th World Wide Web Conference, LNCS 9792*, 2016.
11. M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING '92*, page 539–545, 1992.
12. Amaç Herdağdelen and Marco Baroni. Bagpack: A general framework to represent semantic relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, 2009.
13. J. R. Hobbs and E. Riloff. Information extraction. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.
14. M-C. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *LREC '06*, 2006.
15. M-C. de Marneffe and C. D. Manning. Stanford typed dependencies manual. The Stanford NLP Group, 2008. revised for the Stanford Parser v. 3.7.0 in September 2016.
16. M-C. de Marneffe and C. D. Manning. The stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, 2008.
17. R. J. Mooney and R. Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10, June 2005.
18. Y. Nakamura-Delloye and E. de la Clergerie. Exploitation de résultats d'analyse syntaxique pour extraction semi-supervisée des chemins de relations. In *17e Conférence sur le Traitement Automatique des Langues Naturelles - TALN 2010*, 2010.
19. M. Porumb, I. Barbantan, C. Lemnaru, and R. Potolea. Remed: Automatic relation extraction from medical documents. In *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, iiWAS '15. ACM, 2015.

20. R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *EDBT*, pages 3–17, 1996.
21. P. D. Turney. Measuring semantic similarity by latent relational analysis. In *IJCAI-05*, 2005.
22. P. D. Turney. Similarity of semantic relations. *CoRR*, abs/cs/0608100, 2006.
23. P. D. Turney. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44, 2012.
24. P. D. Turney and S. M. Mohammad. Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering*, 2014.
25. J. Weeds, D. Clarke, J. Reffin, D. Weir, and B. Keller. Learning to distinguish hypernyms and co-hyponyms. In *COLING '14*, 2014.
26. R. Yangarber, W. Lin, and R. Grishman. Unsupervised learning of generalized names. In *COLING '02*, 2002.
27. Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *CIKM*, 2002.
28. Y. Zhao, G. Karypis, and U. Fayyad. Hierarchical clustering algorithms for document datasets. *Data Mining for Knowledge Discovery*, 10, March 2005.
29. G. Zhou, J. Su, J. Zhang, and M. Zhang. Exploring various knowledge in relation extraction. In *ACL '05*, 2005.