Joining Dictionaries and Word Embeddings for Ontology Induction

D.A. Ustalov <dau@imm.uran.ru>
Krasovskii Institute of Mathematics and Mechanics
16 Sofia Kovalevskaya Str., 620990 Yekaterinburg, Russia

Abstract. This paper presents an ontology induction approach that extracts the structured lexical knowledge from synonym dictionaries and establishing the semantic relations within these structures using word embeddings and their projections. The results of the preliminary experiments have also been reported showing certain strengths and weaknesses of the proposed approach.

Keywords: ontology induction; lexical resource; synonyms; word embeddings; graph clustering; projection learning.

DOI: 10.15514/ISPRAS-2016-28(6)-14

For citation: Ustalov D.A. Joining Dictionaries and Word Embeddings for Ontology Induction. Trudy ISP RAN/Proc. ISP RAS, vol. 28, issue 6, 2016, pp. 197-206. DOI: 10.15514/ISPRAS-2016-28(6)-14

1. Introduction

A thesaurus, or a lexical ontology, is a lexical database that groups the words into the sets of synonyms called *synsets* or *concepts*, and records a number of semantic *relations* between these concepts [1]. It is a crucial resource for many natural language processing and artificial intelligence problems, which require common sense reasoning. However, the current state of the electronic thesauri for the Russian language makes it highly topical to refine and to integrate the existing openly available resources to facilitate the development of language technology [2]. In this study, an ontology induction approach that integrates the knowledge represented by the synonym dictionaries enhanced by the methods of distributional semantics has been presented and preliminarily evaluated.

2. Related Work

Given the fact that thesauri are composed of concepts and relations, the approaches for acquisition of both have been briefly reviewed. Unsupervised methods for concept

discovery are designed for clustering co-occurrence graphs to group the words having similar meanings. For instance, the methods proposed by Schütze [3] and Lin & Pantel [4] construct such graphs using large text corpora. Recently, a significant attention in lexical semantics has been paid to the specialized algorithms like Chinese Whispers [5], which is a hard clustering algorithm that assigns a word to at most one cluster at a time, and MaxMax [6], which is a soft clustering algorithm designed specifically for the word sense induction task [7].

Currently, the most widely used method for detecting hyponymy-hypernymy relations is the Hearst patterns [8]. However, these lexical-syntactic patterns offer the sparse representation of words that is less convenient than word embeddings [9]. Fu et al. [10] proposed the projection learning approach to learning hypernyms for the Chinese language. This approach assumes learning the projection matrix so that multiplying it on a hyponym vector produces a hypernym vector. The learning problem has been posed as the linear regression problem that has been then numerically approximated using stochastic gradient descent. Shwartz et al. developed an integrated method that combines the syntactic parsing features with word embeddings based on a long short-term memory network [11]. The resulting method called HypeNET has been implemented using the recurrent neural network that encodes the patterns with the embeddings.

3. Approach

The proposed ontology induction approach, depicted at Fig. 1, uses a synonym dictionary to produce the concepts, and a pre-trained word embeddings model to establish the hyponymy-hypernymy relations between the concepts. The goal of the concept discovery step is to yield a set of concepts by grouping the words with similar meanings from a synonym graph. The goal of the relation establishment step is to link these concepts to each other using the hyponymy-hypernymy relation, which is also known as the subsumption relation or *is-a* relation.

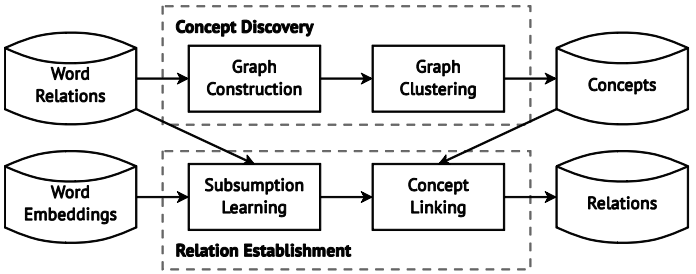

Fig. 1. The proposed ontology induction approach

3.1 Concept Discovery

A synonym dictionary is a graph the set of which vertices contains the words and the set of which edges contains the word pairs connected via the synonymy relation. The cliques in such a graph naturally form the densely connected sets of synonyms [12]

corresponding to the concepts. Given the fact that the clique problem in a graph is NP-complete [13], the Chinese Whispers [5] graph clustering algorithm has been used for finding a global segmentation of the graph. However, the hard clustering property of this algorithm does not handle the polysemous words well. To deal with that, a word sense induction procedure has been run to extract the word senses [14] which have been then combined back into a disambiguated word sense graph [15]. Finally, the disambiguated word sense graph has been clustered again using Chinese Whispers to induce the concepts from these disambiguated word senses.

3.2 Relation Establishment

To extend the lexical coverage of the available subsumption word pairs, a projection learning setup has been used [10] to compute a transformation matrix for word embeddings by projecting the hyponym vector to its possible hypernym vector. To get this done, the 100-dimensional word embeddings dataset for Russian trained by Arefyev et al. [16] using the skip-gram architecture with the sliding window of 10 words and the minimal word frequency of 100 on a text corpus of 13 billion words have been obtained (this configuration is among the best of those participated in the RUSSE study [17] despite its low computational performance requirements). The subsumption pairs from the Russian Wiktionary [18] stem both the train and test sets for learning the projection matrix, the total number of the pairs being 33 885. To avoid lexical overfitting [19], no hyponym from the train set is present in the test set. Since the specificity of the semantic relations differing in various regions of the embedding space, the hyponym embeddings have been clustered using the *k*-means algorithm tuned on a development dataset.

Having the concepts induced and the subsumptions trained, the relations between the concepts are established as follows. For each concept, every word has been projected to its hypernym embedding and the ten nearest neighbours of that projection have been obtained. These neighbours jointly form a bag of words for this concept, for which the most similar concept is computed using the cardinality of the set intersection as the similarity measure. If such a concept found, the former concept is considered as a hyponym of the latter.

4. Experiments

For evaluation purposes, the Russian Wiktionary [18], the Abramov's dictionary [20], and the Universal Dictionary of Concepts [21] have been combined into the single graph to benefit from different lexical coverage provided by the different synonym dictionaries and also to enforce the jointly observed synonymy relations. The resulting graph has 406 889 edges connecting 74 133 individual words. The RuThes-lite 2.0 lexical ontology, composed of 31.5K concepts, 111.5K lexemes and 130K relations, has been used as the gold standard [1] during these experiments.

4.1 Concept Evaluation

To assess the performance of the described concept discovery method, the same graph has been also processed by two other algorithms: Chinese Whispers [5] and MaxMax [6]. Similarly to the experimental setup used for evaluating the MaxMax clustering algorithm [6], the pairwise precision, recall and F_1-measure scores [22] and V-measure score [23] have been computed (Table 1).

Table 1. Concept evaluation

Method \ Parameter	# of lexemes	# of concepts	Precision	Recall	F_1-measure	V-measure
Chinese Whispers	73 878	16 063	0.135	0.022	0.038	0.866
MaxMax	73 878	16 870	0.181	0.004	0.007	0.835
Concept Discovery	29 650	5 984	0.193	0.039	0.065	0.860

According to the concept evaluation results, the described concept discovery method outperformed other methods on every pairwise score and showed the comparable V-measure representing the goodness of the output clustering. As a hard clustering algorithm, Chinese Whispers demonstrated good performance on grouping monosemous words into the concepts like {*компьютер* (computer), *ЭВМ* (ECM), …}, especially on named entities. Also, as anticipated, its performance degraded on polysemous words, resulting in the concepts like {*вода* (water), *акватория* (water area), *влага* (moisture), *кислород* (oxygen), *водород* (hydrogen), …}. Surprisingly, MaxMax, despite the existence of a successful case study of the Portuguese language [7], showed poor results in this study due to the possible difference of the expected graph structure. Firstly, unlike other methods being compared, it emitted a large number of the concepts grouping more than 300 words. After the investigation, these concepts have been removed from the evaluation as the non-relevant. Secondly, a substantial part of the concepts provided by MaxMax grouped the words having no obvious synonymy relation, e.g., {*прайс* (price), *бином Ньютона* (Newtonian binomial), *программный пакет* (software package), …}. In contrast, the concepts yielded by the described concept discovery method correctly reflect the polysemy phenomenon, e.g., {*пустота* (emptiness), *бессодержательность* (barrenness), *бессмысленность* (meaninglessness), …} and {*вакуум* (vacuum), *пустота* (emptiness), *ничто* (nihil), …}. Unfortunately, the disambiguation procedure being used [15] tends to miss certain underrepresented word senses, which results in their absence in the disambiguated word sense graph, and, therefore, in the output concepts as well.

4.2 Relation Evaluation

To assess the performance of the described relation establishment method, it has been applied for each of the 5 984 concepts discovered at the previous step. Each concept has been matched to the most similar RuThes-lite concept using the cardinality of the set intersection as the similarity measure. An established relation has been then considered as correct if there exists a directed path from the hyponym concept to the hypernym concept in the gold standard. Also, the performance of the projection learning setup has been compared to the using of the unmodified subsumption pairs from the Russian Wiktionary without the word embeddings. Table 2 shows the relation evaluation results.

Table 2. Relation evaluation

Dataset \ Parameter	**# of candidate relations**	**# of matched concepts**	**# of correct relations**
Russian Wiktionary	1 627	1 210	113
Projection Learning	3 918	2 293	133

According to the relation evaluation results, the projection learning setup increases the number of candidate concepts and relations, but the number of the correctly established relations did not increase substantially. However, due to the lack of the available subsumption dictionaries, it seems reasonable to try a larger word embeddings dataset, a different learning setup, or a different concept similarity measure.

5. Conclusion

In this short paper, an ontology induction approach that induces a thesaurus structure by integrating synonym dictionaries for discovering concepts and a distributional model for establishing the hyponymy-hypernymy relations between them, has been described and preliminarily evaluated. The results of this study are openly available under a libré license: https://github.com/dustalov/concept-discovery. The plans for the further study include the improving the relation establishment step by more sophisticated matching and machine learning techniques as well as applying crowdsourcing for validating the subsumption candidates.

Acknowledgements. The author is grateful to Andrew Krizhanovsky for the provided Russian Wiktionary dump, to Natalia Loukachevitch for the provided RuThes-lite dataset, and to Alexander Panchenko for the fruitful discussions on the present study. The reported study was funded by RFBR according to the research project no. 16-37-00354 мол_а. This work is supported by the Russian Foundation for the Humanities

projects no. 13-04-12020 ``New Open Electronic Thesaurus for Russian" and no. 16-04-12019 ``RussNet and YARN thesauri integration".

References

[1]. Loukachevitch, N. Thesauri in Information Retrieval Tasks. Moscow State University Press, Moscow, 2011 (in Russian)

[2]. Kiselev, Y., Porshnev, S.V., Mukhin, M. Current Status of Russian Electronic Thesauri: Quality, Completeness and Availability. Programmnaya Ingeneria, 6, 2015, pp. 34–40 (in Russian)

[3]. Schütze, H. Automatic Word Sense Discrimination. Journal of Computational Linguistics, 24, 1998 pp. 97–123

[4]. Lin, D., Pantel, P. Concept Discovery from Text. In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, Association for Computational Linguistics, 2002, pp. 1–7

[5]. Biemann, C. Chinese Whispers: An Efficient Graph Clustering Algorithm and Its Application to Natural Language Processing Problems. In Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing. TextGraphs-1, Association for Computational Linguistics, 2006 pp. 73–80

[6]. Hope, D., Keller, B.: MaxMax: A Graph-Based Soft Clustering Algorithm Applied to Word Sense Induction. In: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I, Springer Berlin Heidelberg, 2013, pp. 368–381

[7]. Gonçalo Oliveira, H., Gomes, P. ECO and Onto.PT: a flexible approach for creating a Portuguese wordnet automatically. Language Resources and Evaluation, v. 48, № 2, 2014, pp. 373–393

[8]. Hearst, M.A. Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th Conference on Computational Linguistics - Volume 2. COLING '92, Association for Computational Linguistics, 1992, pp. 539–545

[9]. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013 pp. 3111–3119

[10]. Fu, R., Guo, J., Qin, B., Che, W., Wang, H., Liu, T. Learning Semantic Hierarchies via Word Embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2014 pp. 1199–1209

[11]. Shwartz, V., Goldberg, Y., Dagan, I. Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2016, pp. 2389–2398

[12]. Kamps, J., Marx, M., Mokken, R.J., de Rijke, M. Using WordNet to Measure Semantic Orientations of Adjectives. In Proceedings of LREC'2004. European Language Resources Association, 2004 pp. 1115–1118

[13]. Bomze, I.M., Budinich, M., Pardalos, P. M., Pelillo, M. The maximum clique problem. In Handbook of Combinatorial Optimization. Springer, 1999, 1–74

[14]. Panchenko, A., Simon, J., Riedl, M., Biemann, C. Noun Sense Induction and Disambiguation using Graph-Based Distributional Semantics. In Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). Bochumer Linguistische Arbeitsberichte, 2016 pp. 192–202

[15]. Faralli, S., Panchenko, A., Biemann, C., Ponzetto, S.P. Linked Disambiguated Semantic Networks. In 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II, Springer International Publishing, 2016 pp. 56–64

[16]. Arefyev, N., Panchenko, A., Lukanin, A., Lesota, O., Romanov, P.: Evaluating Three Corpus-Based Semantic Similarity Systems for Russian. In Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Volume 2 of 2, Papers from special sessions, RGGU, 2015, pp. 106–118

[17]. Panchenko, A., Loukachevitch, N.V., Ustalov, D., Paperno, D., Meyer, C.M., Konstantinova, N.: RUSSE: The First Workshop on Russian Semantic Similarity. In Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Volume 2 of 2, Papers from special sessions, RGGU, 2015, pp. 89–105

[18]. Krizhanovsky, A.A., Smirnov, A.V.: An approach to automated construction of a general-purpose lexical ontology based on Wiktionary. Journal of Computer and Systems Sciences International, v. 52, № 2, 2013, pp. 215–225

[19]. Levy, O., Remus, S., Biemann, C., Dagan, I.: Do Supervised Distributional Methods Really Learn Lexical Inference Relations? In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2015, pp. 970–976

[20]. Abramov, N. The dictionary of Russian synonyms and semantically related expressions, 7th edition. Moscow (1999) (in Russian)

[21]. Dikonov, V.G. Development of lexical basis for the Universal Dictionary of UNL Concepts. In Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Bekasovo, Russia, 2013, pp. 212–221

[22]. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, v. 2, № 4, 2011, pp. 37–63

[23]. Rosenberg, A., Hirschberg, J.: V-Measure. A Conditional Entropy-Based External Cluster Evaluation Measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics, 2007, pp. 410–420

Применение словарей и векторов слов для автоматического построения лексической онтологии

Д.А. Усталов <dau@imm.uran.ru>
Институт математики и механики им. Н.Н.Красовского Уральского отделения Российской академии наук,
Россия, 620990, г. Екатеринбург, ул. Софьи Ковалевской, д. 16

Аннотация. В статье представлен подход к автоматическому построению лексической онтологии путём извлечения и связывания структурированных данных, направленный на повторное использование материалов существующих лексических ресурсов неизвестного качества. Подход состоит из двух этапов. На первом этапе производится построение и кластеризация графа синонимов с целью вывода отдельных значений слов и их объединения в синонимические ряды, именуемые синсетами или понятиями. На втором этапе производится формирование родо-видовых отношений между понятиями путём сопоставления родо-видовых пар слов. С целью расширения множества доступных родо-видовых пар слов выполняется преобразование векторных представлений гипонимов в векторные представления гиперонимов при помощи проекционной матрицы. Проведены предварительные эксперименты с использованием тезауруса русского языка в качестве золотого стандарта. Проанализированы преимущества и недостатки предложенного подхода.

Ключевые слова: автоматическое построение онтологии; лексический ресурс; синоним; векторное представление слова в пространстве низкой размерности;, кластеризация графа; подбор проекционной матрицы.

DOI: 10.15514/ISPRAS-2016-28(6)-14

Для цитирования: Усталов Д.А. Применение словарей и векторов слов для автоматического построения лексической онтологии. Труды ИСП РАН, том 28, вып. 6, 2016 г., стр. 197-206 (на английском). DOI: 10.15514/ISPRAS-2016-28(6)-14

Благодарности. Автор благодарит Андрея Крижановского за предоставленный машиночитаемый вариант Русского Викисловаря, Наталью Лукашевич за предоставленный машиночитаемый вариант тезауруса РуТез, а также Александра Панченко за ценные рекомендации по настоящей работе. Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-37-00354 мол_а. Исследование поддержано грантами РГНФ № 13-04-12020 «Новый открытый электронный тезаурус русского языка» и № 16-04-12019 «Интеграция тезаурусов RussNet и YARN».

Список литературы

[1]. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М.: Издательство Московского университета, 2011. 512 с.

[2]. Киселёв Ю., Поршнев, С.В., Мухин М.Ю. Современное состояние электронных тезаурусов русского языка: качество, полнота и доступность. Программная инженерия, 2015, вып. 6, с. 34-40.

[3]. Schütze H. Automatic Word Sense Discrimination. Journal of Computational Linguistics, 1998, issue 24, pp. 97-123.

[4]. Lin D., Pantel P. Concept Discovery from Text. Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, pp. 1-7, 2002, Association for Computational Linguistics.

[5]. Biemann C. Chinese Whispers: An Efficient Graph Clustering Algorithm and Its Application to Natural Language Processing Problems. Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing (TextGraphs-1), pp. 73-80, 2006, Association for Computational Linguistics.

[6]. Hope D., Keller B. MaxMax: A Graph-Based Soft Clustering Algorithm Applied to Word Sense Induction. Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Part I, pp. 368-381, 2013, Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-37247-6_30.

[7]. Gonçalo Oliveira H., Gomes P. ECO and Onto.PT: a flexible approach for creating a Portuguese wordnet automatically. Language Resources and Evaluation, 2014, vol. 48, issue 2, pp. 373-393. DOI: 10.1007/s10579-013-9249-9.

[8]. Hearst M.A. Automatic Acquisition of Hyponyms from Large Text Corpora. Proceedings of the 14th Conference on Computational Linguistics - Volume 2, pp. 539-545, 1992, Association for Computational Linguistics. DOI: 10.3115/992133.992154.

[9]. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26, pp. 3111-3119, 2013, Curran Associates, Inc.

[10]. Fu R., Guo J., Qin B., Che W., Wang H., Liu T. Learning Semantic Hierarchies via Word Embeddings. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1199-1209, 2014, Association for Computational Linguistics.

[11]. Shwartz V., Goldberg Y., Dagan I. Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2389-2398, 2016, Association for Computational Linguistics.

[12]. Kamps J., Marx M., Mokken R.J., de Rijke M. Using WordNet to Measure Semantic Orientations of Adjectives. Proceedings of LREC'2004, pp. 1115-1118, 2004, European Language Resources Association.

[13]. Bomze I.M., Budinich M., Pardalos P. M., Pelillo M. The Maximum Clique Problem. Handbook of Combinatorial Optimization, 1999, pp. 1-74. DOI: 10.1007/978-1-4757-3023-4_1.

[14]. Panchenko A., Simon J., Riedl M., Biemann C. Noun Sense Induction and Disambiguation using Graph-Based Distributional Semantics. Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), pp. 192-202, 2016, Bochumer Linguistische Arbeitsberichte.

[15]. Faralli S., Panchenko A., Biemann C., Ponzetto S.P. Linked Disambiguated Semantic Networks. Proceedings of the 15th International Semantic Web Conference - Part II, pp. 56-64, 2016, Springer International Publishing. DOI: 10.1007/978-3-319-46547-0_7.

[16]. Арефьев Н., Панченко А., Луканин А., Лесота О., Романов, П. Сравнение трёх систем семантической близости для русского языка. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» - Том 2. Доклады специальных секций, с. 106-118, 2015, Изд-во РГГУ.

[17]. Панченко А., Лукашевич Н.В., Усталов Д., Паперно Д., Мейер К.М., Константинова Н. RUSSE: семинар по оценке семантической близости для русского языка. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» - Том 2. Доклады специальных секций, с. 89-105, 2015, Изд-во РГГУ.

[18]. Krizhanovsky A.A., Smirnov A.V. An approach to automated construction of a general-purpose lexical ontology based on Wiktionary. Journal of Computer and Systems Sciences International, 2013, vol. 52, issue 2, pp. 215-225. DOI: 10.1134/S1064230713020068.

[19]. Levy O., Remus S., Biemann C., Dagan I. Do Supervised Distributional Methods Really Learn Lexical Inference Relations?. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 970-976, 2015, Association for Computational Linguistics.

[20]. Абрамов Н. Словарь русских синонимов и сходных по смыслу выражений. М.: Русские словари, 1999, 528 с.

[21]. Dikonov V.G. Development of lexical basis for the Universal Dictionary of UNL Concepts. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, pp. 212-221, 2013. RGGU.

[22]. Powers D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2011, vol. 2, issue 4, pp. 37-63.

[23]. Rosenberg A., Hirschberg J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 410-420, 2007, Association for Computational Linguistics.