Unsupervised Construction of a Multilingual WordNet from Parallel Corpora

Dimitar Kazakov and Ahmad R. Shahid
Department of Computer Science
University of York
Heslington, York YO10 5DD, UK
kazakov|ahmad@cs.york.ac.uk

Abstract

This paper outlines an approach to the unsupervised construction from unannotated parallel corpora of a lexical semantic resource akin to WordNet. The paper also describes how this resource can be used to add lexical semantic tags to the text corpus at hand. Finally, we discuss the possibility to add some of the predicates typical for WordNet to its automatically constructed multilingual version, and the ways in which the success of this approach can be measured.

Keywords

Parallel corpora, WordNet, unsupervised learning

1 Introduction

Lexical ambiguity is inherent and widespread in all languages; it emerges spontaneously in computer simulations of language evolution [23], and its origin probably stems in the interplay between geographic divisions and interaction between communities, diachronic linguistic changes, and evolutionary pressures on the cost of communication. Two challenges arise when dealing with lexical ambiguity: firstly, to define the elementary semantic concepts employed in a given language, and, secondly, given one or more utterances, to identify the semantic concepts to which the words in those utterances refer. Throughout history, numerous attempts have been made to address both challenges through the construction of artificial languages with unambiguous semantics (*e.g.*, see Ecos detailed and entertaining review [3]). Arguably, there are two possible ways of defining a semantic concept, either by trying to map it onto some sensory experience (leading to a philosophical discussion about the extent to which they are shared and the notion of *qualia*), or by describing it through other concepts, in a way that is mutually recursive and unbounded (cf. Peirces Sign Theory and the notion of *infinite semiosis*).

The last twenty five years saw academic and commercial efforts directed towards the creation of large repositories combining the description of semantic concepts, their relationship, and, possibly, common knowledge statements expressed in those terms. This includes, among others, the Cycorp Cyc project [11] and the lexical semantic database WordNet [14]. Both approaches use a number of predicates to make statements, such as "concept A is an instance of concept B" or "concept A is a specific case of concept B" (in other terms, all instances of concept A form a subset of the instances of concept B). WordNet also employs the notion of *synsets*, defining a semantic concept through the list of words (synonyms) sharing that meaning, *e.g.*, {mercury, quicksilver, Hg}. The original version of WordNet developed in Princeton covered the English language, but this effort has been replicated for other languages [25]. All these databases are monolingual; they are also handcrafted, and while very comprehensive in many aspects, it is difficult to assume they could be kept abreast of the new developments, including the use of newly coined words, and giving new meanings to the old ones.

The dawn and rapid growth of dynamic online encyclopedic resources with shared authorship, such as Wikipedia, in the last decade, have begun to draw attention as a potential source of semantic concepts and ontologies [7]. Recently, there have also been efforts to use the likes of Wikipedia to the existing hand-crafted resources [13].

2 Multilingual Synsets

The synsets in WordNet clarify a concept (or, from another point of view, narrow down the sense of a word) by paraphrasing it repeatedly, using other words of the same language. This approach is based on the fact that words rarely share all their meanings: {step, pace} specifies a different meaning from {step, stair}. The same result, however, can be achieved through the use of words of different languages that share at least one sense and therefore can be seen as translations of each other. So, {EN: bank, FR: banque} could refer to a financial institution or a collection of a particular kind (*e.g.*, a blood bank), as these words share both meanings, but eliminates the concept of 'river bank' that the English word alone might denote. Increasing the number of languages could gradually remove all ambiguity, as in the case of {EN: bank, FR: banque, NL: bank}. Insofar these lists of words specify a single semantic concept, they can be seen as synsets of a WordNet that makes use of words of several languages, rather than just one. The greater the number of translations in this multilingual WordNet, the clearer the meaning, yet, one might object, the fewer the number

NLP Methods and Corpora in Translation, Lexicography, and Language Learning 2009 – Borovets, Bulgaria, pages 9–12

of such polyglots, who could benefit from such translations. However, these multilingual synsets can also be useful in a monolingual context, as unique indices that distinguish the individual meanings of a word. For instance, if the English word bank is translated variously as {EN: bank, FR: banque}, and {EN: bank, FR: rive} one does not need to understand French to suspect that 'bank' may have (at least) two different meanings. The greater the number of languages in which the two synsets differ, the stronger the intuition that they indicate different meanings of the word in the pivotal language.

Synsets, whether monolingual or multilingual, can be used to tag the lexical items in a corpus with their intended meaning (see Table 1). The benefits of such lexical semantic tags are evident. Focussing now on the novel case of multilingual synsets, one can consider the two separate questions of how to produce a collection of such synsets, and how to annotate the lexical items in a text with them. Kazakov and Shahid [22] present an interesting study in that direction, where the titles of 'equivalent' Wikipedia pages are collected for a number of preselected languages on the assumption that they represent an accurate translation of each other (see Fig.1).

The authors repeat the same exercise for a number of specific domains, and also with the names of Wikipedia categories. The latter case demonstrates the potential to use Wikipedia not only to collect multilingual synsets, but also to add a hierarchical relationship between them, as this information is explicitly present there. A number of other researchers have in fact used Wikipedia to extract ontologies [6], [7], but in all cases this was done for a single language.

Semantically disambiguated corpora, including those using WordNet synsets as semantic tags, are valuable assets [10], [9], but creating them requires a considerable effort. Here we propose an alternative approach, where a text is automatically annotated with lexical semantic tags in the form of multilingual synsets, provided the text forms part of a multilingual, parallel corpus.

Table 1: *Examples of lexical semantic annotation using standard English WordNet synsets (above) and multilingual synsets (below).*

A darkened outlook for Germany's banks:**SS1** The banks:**SS2** of the river Nile **SS1**={bank, depository financial institution} Gloss="a financial institution that accepts deposits and lends money" **SS2**={bank} Gloss="sloping land"
A darkened outlook for Germany's banks:**mSS1** The banks:**mSS2** of the river Nile **mSS1**={EN:bank, FR:banque, CZ:banka} **mSS2**={EN:bank, FR:rive, CZ:břeh}

Table 2: *Examples of variation between synsets due to the use of different word forms (above) and synonyms (below).*

EN	**FR**	**CZ**	...
work	*travail*	práce	...
works	*travaux*	práce	...
work	*travail*	práce	...
work	*bouleau*	práce	...

3 Annotating Parallel Corpora with Lexical Semantics

In our approach the multilingual synsets become the sense tags and the parallel corpora are tagged with the corresponding tags (see Fig.2).

The idea is as simple as it is elegant: assuming we have a word-aligned parallel corpus with n languages, annotate each word with a lexical semantic tag consisting of the n-tuple of aligned words. As a result, all occurrences of a given word in the text for language $\mathcal{L}$ are considered as having the same sense, provided they correspond to (are tagged with) the same multilingual synset.

Two great advantages of this scheme are that it is completely unsupervised, and the fact that, unlike manually tagged corpora using WordNet, all words in the corpus are guaranteed to have a corresponding multilingual synset. Since we are only interested in words with their own semantics, a stop list can be used to remove the words of the closed lexicon before the rest are semantically tagged. Also the need for word alignment should not be an issue, at least in principle, as there are standard tools, such as GIZA++ [16] serving that purpose.

The approach as described so far needs to deal with two main issues, both related to the fact that the simple listing of n-tuples as synsets may produce many more synsets than the real number of concepts to which the words in the text refer. The first issue stems from the fact that a lexeme (word entry) corresponds to several word forms in most languages, so, for instance, the word forms {EN: work} and {EN: works} will correspond to two different synsets (Table 2, top), even if they are used with the same meaning. The second of the above mentioned issues is related to the use of synonyms in one language, whereas the translation in another makes use of the same word (lexeme) (Table 2, bottom).

From this point of view, we could consider the original n-tuples as *proto-synsets*, and then strife to recognize the variation due to the use of different word forms and synonyms, and eliminate it, if possible, by grouping these proto-synsets into genuine synsets corresponding to different concepts. As much of the appeal of the whole approach stems from its unsupervised nature, we shall also consider unsupervised ways of solving this specific problem. For several languages, there are detailed, explicit models of their word morphology [19], [20], [17], which makes mapping word forms onto the list of lexemes they may represent a straightforward task.

English	German	French	Polish	Bulgarian	Greek	Chinese
Wikipedia	Wikipedia	Wikipédia	Wikipedia	Уикипедия	Βικιπαίδεια	維基百科
Encyclopedia	Enzyklopädie	Encyclopédie	Encyklopedia	Енциклопедия	Εγκυκλοπαίδεια	百科全书
English language	Englische Sprache	Anglais	Język angielski	Английски език	Αγγλική γλώσσα	英语
Venice	Venedig	Venise	Wenecja	Венеция	Βενετία	威尼斯
Film director	Regisseur	Réalisateur	Reżyser	Режисьор	Σκηνοθέτης	電影導演
Uniform Resource Locator	Uniform Resource Locator	Uniform Resource Locator	Uniform Resource Locator	Унифициран локатор на ресурси	Uniform Resource Locator	统一资源定位符
Web search engine	Suchmaschine	Moteur de recherche	Wyszukiwarka internetowa	Търсачка	Μηχανή αναζήτησης	搜索引擎
University	Hochschule	Université	Uniwersytet	Университет	Πανεπιστήμιο	大學
Monopoly	Monopol	Monopole	Monopol	Монопол	Μονοπώλιο	垄断
Computer	Computer	Ordinateur	Komputer	Компютър	Ηλεκτρονικός υπολογιστής	計算機
University of Oxford	University of Oxford	Université d'Oxford	Uniwersytet Oksfordzki	Оксфордски университет	Πανεπιστήμιο της Οξφόρδης	牛津大学
Population density	Bevölkerungsdichte	Densité de population	Gęstość zaludnienia	Гъстота на населението	Πυκνότητα πληθυσμού	人口密度
Presidential system	Präsidentielles Regierungssystem	Régime présidentiel	System prezydencki	Президентска република	Προεδρική Δημοκρατία	總統制
Dictatorship	Diktatur	Dictature	Dyktatura	Диктатура	Δικτατορία	专政
European Community	Europäische Gemeinschaft	Communauté européenne	Wspólnota Europejska	Европейска общност	Ευρωπαϊκή Κοινότητα	欧洲共同體
Benazir Bhutto	Benazir Bhutto	Benazir Bhutto	Benazir Bhutto	Беназир Бхуто	Μπεναζίρ Μπούτο	贝娜齐尔·布托
Thomas Edison	Thomas Alva Edison	Thomas Edison	Thomas Alva Edison	Томас Едисън	Τόμας Έντισον	托马斯·爱迪生
Art	Kunst	Art	Sztuka	Изкуство	Τέχνη	艺术
California	Kalifornien	Californie	Kalifornia	Калифорния	Καλιφόρνια	加利福尼亚州
Buddhism	Buddhismus	Bouddhisme	Buddyzm	Будизъм	Βουδισμός	佛教

Fig. 1: *Wikipedia page titles seen as multilingual synsets.*

Using morpho-lexical analyzers for the languages in the corpus will produce a list of lexical entries for each language, which can be narrowed down through the use of conventional lexicons listing the possible pairs of lexical entries between given two languages. In this way, the word form 'works' will be mapped onto the lexemes *work*, **n.**, *works*, **n.**, and *work*, **v.**, but in the context of the French *travail*, only the first lexeme will be retained, as the other two would not form a translation pair in an English-French lexicon.

One could also consider doing away with any models of morphology, assuming complete ignorance in this respect, and employing unsupervised learning of word morphology [8], [4]. In their latest form, these approaches can identify word form paradigms, which would allow all forms of a lexical entry to be mapped consistently onto it.

It is also possible to automate the process of identifying synonyms among the words of a given language. For instance, Crouch's approach [2] is based on the discrimination value model [21]. Other approaches include Bayesian Networks [18], Hierarchical Clustering [24], and link co-occurrences [15], etc. Such automated approaches have certain advantages over the manually generated thesauri, both in terms of cost and time of development, and also in the level of detail, with the latter often being too fine grained, and reflecting usages that are not common and irrelevant in practice [12].

4 Extracting a Fully-Fledged Multilingual WordNet

So far, we have described a procedure that extracts multilingual synsets from a parallel corpus and reduces spurious ambiguity by merging synsets corresponding to multiple word forms of the same lexeme, resp. by combining those only varying in the use of synonyms of a given language. Extraction of hierarchical semantic relationships between words in a corpus has been studied for almost two decades [5], and the same procedures can be applied here, leading to the acquisition of a lexical resource akin to WordNet, which also contains some of the predicates (*e.g.*, hyponym/2, resp. hypernym/2). Such semantic hierarchies can then be used to tag the text corpus with synsets of a certain level of granularity, depending on the task at hand.

5 Evaluation

The evaluation of this approach could be twofold: directly, using an already semantically annotated gold standard, and indirectly, through the measured benefits of lexical semantic tagging in other tasks. The limitations of the direct approach are given by the lack of semantically annotated parallel corpora, however, there is at least one such corpus at the time of writing, namely, multi-Semcor [1]. Indirectly, the potential benefits of tagging text with such multilingual synsets can be measured in tasks, such as document clustering, where the lexical semantic tags can be used to discriminate between various word senses. Any improvement in the within-clusters and between-clusters quality measures would indicate relative (and measurable) success.

References

[1] L. Bentivogli, E. Pianta, and M. Ranieri. Multisemcor: an English Italian aligned corpus with a shared inventory of senses. In *Proceedings of the Meaning Workshop*, page 90, Trento, Italy, February 2005.

[2] C. J. Crouch. A Cluster-Based Approach to Thesaurus Construction. In *Proceedings of ACM SIGIR-88*, page 309320, 1988.

[3] U. Eco. *La recherche de la langue parfaite dans la culture européenne*. Seuil, Paris, 1994.

[4] J. Goldsmith. Unsupervised acquisition of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

[5] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, 1992.

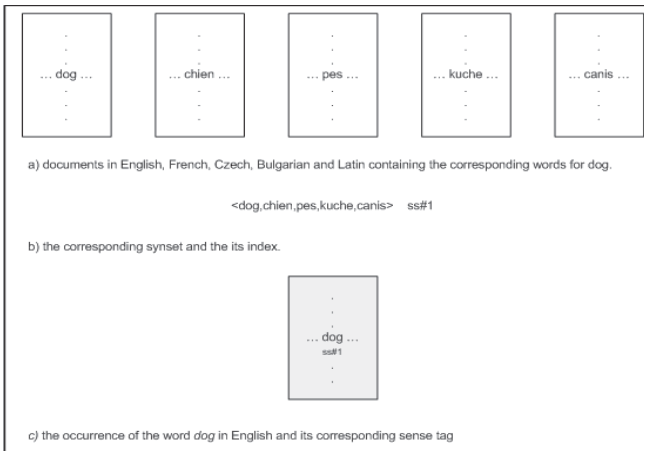

Fig. 2: *Assignment of sense tags in aligned documents.*

[6] M. Hepp, D. Bachlechner, and K. Siorpaes. Harvesting wiki consensus – using wikipedia entries as ontology. In *Proc. ESWC-06 Workshop on Semantic Wikis*, pages 132–46, 2006.

[7] A. Herbelot and A. Copestake. Acquiring ontological relationships from wikipedia using RMRS. In *Proc. ISWC-06 Workshop on Web Content Mining with Human Language*, 2006.

[8] D. Kazakov. Unsupervised learning of nave morphology with genetic algorithms. In W. van den Bosch and A. Weijters, editors, *in Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pages 105–112, Prague, Czech Republic, April 1997.

[9] D. Kazakov. Combining LAPIS and WordNet for the learning of LR parsers with optimal semantic constraints. In S. Dzeroski and P. Flach, editors, *The Ninth International Workshop ILP-99*, volume 1634 of *LNAI*, Bled, Slovenia, 1999. Springer-Verlag.

[10] S. Landes, C. Leacock, and R. Tengi. *WordNet: An Electronic Lexical Database*, chapter Building semantic concordances. MIT Press, Cambridge, Mass., 1998.

[11] D. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), 1995.

[12] D. Lin. Automatic Rretrieval and Clustering of Similar Words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Montreal, Quebec, Canada, August 1998.

[13] O. Medelyan and C. Legg. Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense. In *Proceedings of the WikiAI Workshop at AAAI-2008*, Chicago, 2008.

[14] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five papers on WordNet. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press, May 1998.

[15] K. Nakayama, T. Hara, and S. Nishio. Wikipedia mining for an association web thesaurus construction. In *In Proceedings of IEEE International Conference on Web Information Systems Engineering*, pages 322–334, 2007.

[16] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[17] K. Oflazer. Two-level description of turkish morphology. In *EACL*, 1993.

[18] Y. C. Park and K.-S. Choi. Automatic thesaurus construction using Bayesian networks. *Information Processing and Management: an International Journal*, 32(5):543–553, September 1996.

[19] E. Paskaleva. A formal procedure for Bulgarian word form generation. In *COLING*, pages 217–221, 1982.

[20] G. D. Ritchie, G. J. Russell, A. W. Black, and S. G. Pulman. *Computational Morphology: Practical Mechanism for the English Lexicon*. MIT Press, 1991.

[21] G. Salton, C. Yang, and C. Yu. A Theory of Term Importance in Automatic Text Analysis. *Journal of the American Society for Information Science*, 26(1):33–44, 1975.

[22] A. R. Shahid and D. Kazakov. Automatic Multilingual Lexicon Generation using Wikipedia as a Resource. In *Proc. of the Intl. Conf. on Agents and Artificial Intelligence (ICAART)*, Porto, Portugal, January 2009.

[23] L. Steels. Emergent adaptive lexicons. In P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, and S. W. Wilson, editors, *Fourth International Conference on Simulation of Adaptive Behavior*. The MIT Press/Bradford Books, 1996.

[24] T. Tokunaga, M. Iwayama, and H. Tanaka. Automatic thesaurus construction based-on grammatical relations. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1308–1313, 1995.

[25] P. Vossen, editor. *EuroWordNet*. Kluwer, 1998.