Learning Semantic N-Ary Relations from Wikipedia

Marko Banek, Damir Jurić, and Zoran Skočir

University of Zagreb, Faculty of Electrical Engineering and Computing,
Unska 3, HR-10000 Zagreb, Croatia
{marko.banek,damir.juric,zoran.skocir}@fer.hr

Abstract. Automated construction of ontologies from text corpora, which saves both time and human effort, is a principal condition for realizing the idea of the Semantic Web. However, the recently proposed automated techniques are still limited in the scope of context that can be captured. Moreover, the source corpora generally lack the consensus of ontology users regarding the understanding and interpretation of ontology concepts. In this paper we introduce an unsupervised method for learning domain n-ary relations from Wikipedia articles, thus harvesting the consensus reached by the largest world community engaged in collecting and classifying knowledge. Providing ontologies with n-ary relations instead of the standard binary relations built on the subject-verb-object paradigm results in preserving the initial context of time, space, cause, reason or quantity that otherwise would be lost irreversibly. Our preliminary experiments with a prototype software tool show highly satisfactory results when extracting ternary and quaternary relations, as well as the traditional binary ones.

1 Introduction

Ontologies have been designed as the core of the Semantic web. Early ontologies, constructed manually by domain experts, were of low usability due to their limited scope and to disagreements in understanding and interpreting their content, which arose between the ontology creators and the majority of potential users (who were not involved in the process of ontology creation or were not even able to propose some extensions). In the recent years, with a huge amount of available digital content, methodologies have been developed for automatically creating and populating ontologies from text sources. Most of them focus on class hierarchies and identifying instances, capturing only a minimum of class-to-class relations, such as is-a, part-whole or some relations with generic names such as *associated-with* or *related-to*, which do not represent a satisfactory semantic meaning.

Several existing ontology population techniques able to extract arbitrary semantic relations from text corpora [3,4,8,12] are focused exclusively on binary relations. Being noun-oriented in general, ontologies present binary relations (called properties in OWL [15]) as associations of the pivot domain class with

P. García Bringas et al. (Eds.): DEXA 2010, Part I, LNCS 6261, pp. 470–477, 2010.

other classes. On the contrary, when humans express their thoughts by means of a natural language, they focus on verbs (at least, this is true for Indo-European languages, including English). Consider the following sentence: *John invited his friends to dinner on Friday.* The verb, *invite*, connects four nouns: *John*, *friends*, *dinner* and *Friday.* Certainly, the presented fourfold relation (as seen from the verb as the pivot) can be dissolved into many binary relations to populate ontologies in the standard manner. Recently, extensions to OWL have been proposed to include n-ary relations [14]. We believe that learning the original n-ary relations from the domain can lead to a significant improvement of ontology quality.

While the automatic construction of ontologies from text corpora provides a significant improvement in comparison to manual creation, there still remains an unsolved problem of selecting relevant sources, where there will be no problems with understanding and interpreting concepts. According to [7], ontologies are not just formal representations of a domain, but rather community contracts about such formal representations. Therefore, in order to be able to reflect the community consensus, the vast amount of Wikipedia entries should be reused as ontology components [7]. A Wikipedia entry (i.e. a titled article) may correspond to a class, an instance or a property.

In this paper we introduce an unsupervised method for learning domain n-ary relations from Wikipedia articles. By populating the ontologies with n-ary relations we achieve their higher expressivity and preserve the initial context that otherwise would be lost irreversibly. Another novelty of our approach is the fact that we extract only relations where the participating nouns correspond to Wikipedia titles. Hence, instead of having first to determine which nouns are relevant for the future ontology, which is necessary when relations are extracted from text corpora, we exploit the consensus of Wikipedia article authors regarding term relevance.

The paper is structured as follows. Related work is outlined in Sec. 2. The use of n-ary relations for discovering richer semantic context is explained in Sec. 3. The relation extraction process is elaborated in Sec. 4. The evaluation of the approach is illustrated in Sec. 5. In Sec. 6 conclusions are drawn and a list of future work tasks is presented.

2 Related Work

The initial efforts to extract relations from text, based on pattern matching [11], have worked well for a limited scope of relations which are known to hold in advance (such as *is-a* or *part-of*). The need of the ontology learning community to extract arbitrary relations has produced various methods that first discover verbs in text and then exploit syntactic dependencies between those verbs and other sentence constituents containing nouns [3,4]. In the next step the domain and the range of the relations must be generalized since the nouns in the text are mostly proper nouns or concepts at a level lower than the highest possible domain and/or range concept [4].

An unsupervised method for learning arbitrary semantic binary relations between ontological concepts in the molecular biology domain is presented in [3].

Sentences are parsed using a constituent syntactic parser and then searched in order to find the occurrences of the six relevant syntactic structures. A null hypothesis is stated implying that a particular pair of nouns and a particular verb forming the given syntactic structure do not occur together more frequently than expected at chance. Rejection of the null hypothesis by the subsequently applied chi-square test implies a new relation.

The unsupervised methodology presented in [12] assumes an incremental process, starting with the user's specification of the domain with a single term, and then extracting the relations containing the term, its hyponyms and hypernyms. Heuristic measures related to the occurrence of verbs in the source text are applied to single out the relations that are indeed relevant for the domain.

Meanwhile, the information extraction (IE) community has developed unsupervised methods to extract binary [1] or even n-ary relations [5] from web documents. However, those approaches use documents collected by search engines and apply a less restrictive approach to exploring syntactic dependencies (in comparison with the ontology learning community), which results in a very large portion of "dirty" relations unusable for ontology population.

3 N-Ary Relations in Text

Ontologies can be reused for multiple purposes and by a variety of users if they are able to capture precisely the context of the domain i.e. the context of the sentences that serve as their source. In English, the two main parts of a sentence are the subject and the predicate. The predicate must contain a verb, and, depending on the verb, may also contain objects (direct, indirect and prepositional), predicatives and/or adverbials. Relations capture the association between the verb and the nouns in the sentence, which are either classes or individuals (class instances). Binary relations describe an association between the subject noun, the verb and a noun in the predicate. On the other hand, n-ary relations give much richer and subtler definition of the context: in the majority of cases they include subject, object and prepositional phrases containing nouns that describe the context of time, space, etc.

We will show the plenty of context that can be inferred from a single sentence (taken from the Wikipedia article about Alexander the Great): *Born in Pella in 356 BC, Alexander succeeded his father Philip II of Macedon to the throne in 336 BC, after the King was assassinated, and died thirteen years later at the age of 32*. The main verb, *succeeded*, describes an event of *succession* (WordNet [6] enables finding derivationally related nouns of input verbs). A ternary relation (Fig. 1) associates the event to Alexander (subject), Philip (object) and year 336 BC (prepositional phrase). Particular associations obtain their names from the verb by adding suffices corresponding to the part of the sentence (subject: *is_successor*, object: *succeeded_by*). Possible inverse associations (e.g. *precede-succeed*) can be named by retrieving verb antonyms from WordNet [6].

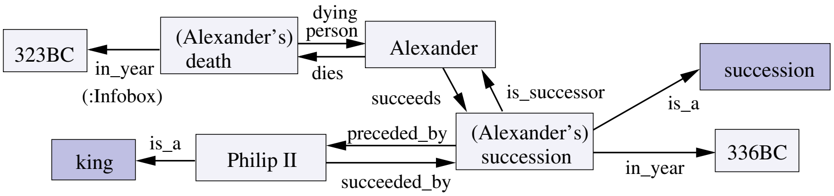

Fig. 1. Ontology structure extracted from Wikipedia article

4 Processing Wikipedia

As mentioned in Sec. 1, we create ontologies with classes and instances corresponding to Wikipedia titles. Hence, when we extract relations from sentences, all nouns declared as relation participants must be annotated as Wikipedia titles.

The source for the entire process are text files formatted in a special kind of markup. We shortly outline its features that are relevant for the relation extraction process. Each article has a title that exactly corresponds to its URL. A sentence from the article titled Napoleon I of France is given in Fig. 2. Links to other articles are given in double square brackets. The displayed text seeding the link may be different from the link title, which is denoted by a vertical bar within the double square brackets (`[[Battle of Austerlitz |Austerlitz]]` in Fig. 2 means a link to the article Battle of Austerlitz, which is simply displayed as `Austerlitz`). The lack of a vertical bar means that the displayed text is identical to the title of the linked article (e.g. `[[Russia]]` in Fig. 2). It is also very important to note that links to other articles appear only once or several times within the text, although the denoted term itself may appear many times (there is a link to Russia, but no link to Austria, since the latter already appears earlier in the text).

> Six weeks later, on the first anniversary of his coronation, Napoleon defeated Austria and [[Russia]] at [[Battle of Austerlitz | Austerlitz]], ending the third coalition.

Fig. 2. A portion of Wikipedia article Napoleon I of France

4.1 Processing Algorithm

Our algorithm for learning n-ary relations from Wikipedia articles consists of several activities (Fig. 3), which will be described in this subsection.

Title Propagation. In order to extract relations from Wikipedia text, we have to identify all occurrences of word tokens that are unambiguously related to the title of a Wikipedia article. Certainly, each link (i.e. double square brackets) points to a title. However, one of the conventions on writing Wikipedia articles is to create a link only for the first occurrence of titled term, while giving no

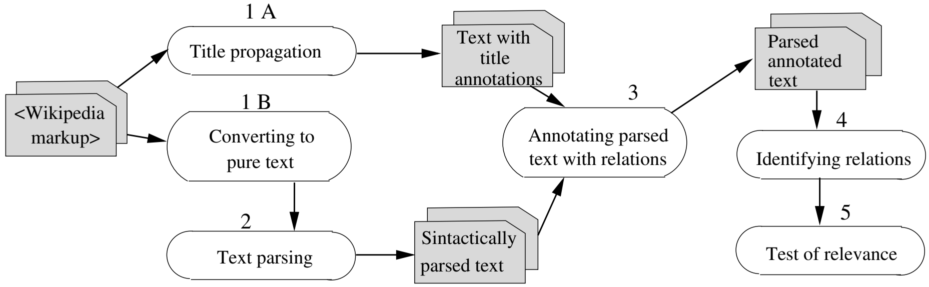

Fig. 3. The relation extraction algorithm

markup emphasis for any further appearances. Hence, for each title, which is either a link indicated by double square brackets or the title of the article itself, we try to find its other occurrences within the article. We call this procedure title propagation. The output of title propagation for the sentence in Fig. 2 is the following (capped text denotes titles): *Six weeks later, on the first anniversary of his coronation,* Napoleon I of France *defeated* Austria *and* Russia *at* Battle of Austerlitz*, ending the third coalition.*

For each title consisting of N multiple words we produce all possible subsets sized from 1 to $N-1$ where each word is allowed to follow only its immediate predecessor in the original title. Subsets that do not contain at least one noun or those that start or end with a preposition or a conjunction are eliminated (Fig. 4). Some subsets may be derived from more than one title present in the article. For instance, titles Louis XVI of France and Louis XVIII of France both produce the subset {Louis}. Consequently, we will not associate {Louis} with any title. The shorter the subset, the more ambiguity is to be expected. Therefore, we start the process of relating text tokens to titles with the longest subsets (all examples of {Louis, XVI} will be associated to Louis XVI of France before even starting with the ambiguous subset {Louis}).

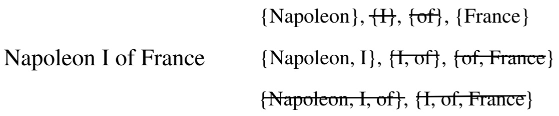
Napoleon I of France
{Napoleon}, ~~{I}~~, ~~{of}~~, {France}
{Napoleon, I}, ~~{I, of}~~, ~~{of, France}~~
~~{Napoleon, I, of}~~, ~~{I, of, France}~~

Fig. 4. Creating subsets during the title propagation process

We do perform title association if the ambiguous subset corresponds to the titular of the article. All occurrences of {Napoleon} in the Napoleon I of France article are associated with Napoleon I, although one actually refers to his nephew, Napoleon III of France. Hence, there may be false title associations, but we argue that producing several false associations is a reasonable price in comparison with losing many relations having the article titular as participant.

A related problem, linking portions of unstructured text to Wikipedia articles, is described in [10]. However, the approaches cannot be compared since their goals require substantially different behavior in case of term ambiguity.

Parsing Sentences. For each input sentence the Stanford dependency parser [9] produces its typed dependency graph, which describes all sentence relationships uniformly as typed relations. Parsing the sentence in Fig. 2 produces the output shown in Fig. 5. There are 55 relation types in total, but we ignore all clause relations (e.g. clausal complement, *xcomp*) and general dependencies (*dep*). The most important relations of our interest are active and passive subject (*nsubj*, *nsubjpass*), passive verb agent (*agent*), direct, indirect or prepositional object (*dobj*, *iobj*, *pobj*) and prepositional modifier (*prep*). Graphs containing conjunction (*conj*) may be split into several independent subgraphs.

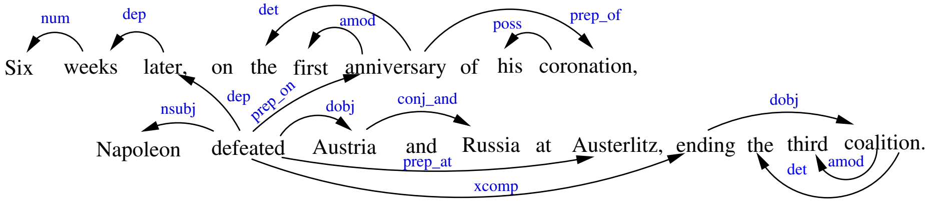
num
dep
det
amod
poss
prep_of
Six weeks later, on the first anniversary of his coronation,
dep
prep_on
nsubj
dobj
conj_and
dobj
Napoleon defeated Austria and Russia at Austerlitz, ending the third coalition.
prep_at
det
amod
xcomp

Fig. 5. The typed dependency graph produced by the Stanford dependency parser

Identifying Relations and Relevance Test. An n-ary relation stemming from a verb v is a set of triples $\{(C, v, r)\}$, where C is an ontology concept (in our case a Wikipedia title), and r is a relation type. All triples share the same verb v and none of them share the same relation type. Considering the output of title propagation parsing, the sentence shown in Fig. 5, produces two ternary relations: { (Napoleon I of France, defeat, *nsubj*), (Austria, defeat, *dobj*), (Battle of Austerlitz, defeat, *prep_at*) } and { (Napoleon I of France, defeat, *nsubj*), (Russia, defeat, *dobj*), (Battle of Austerlitz, defeat, prep_at) }. We use two different methods to identify relevant relations among a much larger number of the automatically identified ones: the chi-square test and a heuristic probabilistic formula. Wikipedia titles within the relevant relations can further be turned into WordNet concepts by following the approach explained in [13].

5 Evaluation

We processed 171 Wikipedia articles describing rulers or military commanders (46 articles), writers (71), philosophers (32) and battles (22). In total, this corpus contained 32809 sentences with 521334 words, which also included 24124 links (4,6% of the total word count). After title propagation, the number of title references increased to 65526 (12,5% of the total word count). They pointed to 15311 different titles, meaning that each of the references appeared 4.3 times in average. We automatically identified binary, ternary and quaternary relations

Table 1. Results of the relevance test

		CHI-SQUARE				AE			
N	II	RR	ST	RR/II	P	RR	ST	RR/II	P
2	1587	99	77	0,062	0,778	981	546	0,618	0,556
3	142	0	0	0	-	100	78	0,704	0,780
4	4	0	0	0	-	3	3	0,75	1

and tested their relevance by using two different approaches: the chi-square test and a heuristic probabilistic formula.

The chi-square test is applied in a fashion similar to [3]. Let A be an ordered tuple of Wikipedia titles participating in the n-ary relation. Let B be a pair, consisting of the verb and an ordered tuple of relation types (the order of relation types corresponds to the order of titles in A). The null hypothesis states that A and B do not co-occur more frequently than by chance. We perform the test using the log-likelihood formula (G^2). Relevant relations produced as the output of the algorithm are those for which the null hypothesis is rejected.

The other test applies the heuristic *above expectation* formula originally devised in [8]. The formula presented in Eq. 1 measures the conditional frequency of all relation concepts (given the verb and the relation types) compared to the conditional frequencies expected when each concept (Wikipedia title in our case) is related to the verb independently of other concepts:

$$AE(C_1 \wedge \ldots \wedge C_n \mid v \wedge r_1 \wedge \ldots \wedge r_n) = \frac{P(C_1 \wedge \ldots \wedge C_n \mid v \wedge r_1 \wedge \ldots \wedge r_n)}{P(C_1 \mid v \wedge r_1) \cdot \ldots \cdot P(C_n \mid v \wedge r_n)} \quad (1)$$

The presented equation is our extension of the original formula [8], which could not distinguish between relation types.

The results of the evaluation are shown in Table 1 (II - no. of initially identified relations; RR - no. of relevant relations; ST - no. of semantically true i.e. correct relations; P - precision i.e. ST/RR). Due to data sparseness, the chi-square test cannot confirm the relevance of any ternary or quaternary relation. However, the AE measure (with a threshold empirically set to 5) filters about 70% of the automatically identified relations with a satisfactory precision of 78%. Although the number of relevant binary relations obtained by the chi-square test is smaller than expected (6.2%), we still prefer chi-square to the AE measure due to higher precision (77,8% in comparison with 55,6%). Thus, we suggest using chi-square test to determine the relevant relations, while the AE measure can serve as its replacement in case of sparse data (n-ary relations with $N > 2$).

6 Conclusion

In this paper we presented an unsupervised approach for learning semantic n-ary relations from Wikipedia. We restrict the nouns participating in relations to those corresponding to Wikipedia article titles, thus harvesting the consensus reached by the authors of Wikipedia - the largest world community engaged in collecting and classifying knowledge. The process of discovering all occurrences of article titles within Wikipedia text, called title propagation, is a key step in identifying the relations. Due to data sparseness, the relevance of the identified relations with more than two members cannot be determined by the chi-square test, otherwise used for binary relations. Instead, satisfactory results are achieved by applying our extension of the "*above expectation*" formula.

Our future work will address the process of naming the extracted relations. Particular focus will be given to discovering verb-preposition patterns that imply the class of the noun in a prepositional phrase: place, time, etc. We will also exploit Wikipedia structures such as infoboxes for acquiring additional semantics.

References

1. Banko, M., Etzioni, O.: The Tradeoffs between Open and Traditional Relation Extraction. Proc. Assoc. Comp. Linguistic (2008)
2. Buitelaar, P., Cimiano, P. (eds.): Ontology Learning and Population: Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text. IOS Press, Amsterdam (2008)
3. Ciaramita, M., Gangemi, A., Ratsch, E., Šarić, J., Rojas, I.: Unsupervised Learning of Semantic Relations for Molecular Biology Ontologies. In: [2]
4. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer, Heidelberg (2006)
5. Eichler, K., Hemsen, H., Neumann, G.: Unsupervised Language Extraction from Web Documents. In: Proc. LREC 2008, pp. 1674–1679 (2008)
6. Fellbaum, C. (ed.): WordNet. An Electronic Lexical Database. MIT Press, Cambridge (1998)
7. Hepp, M., Bachlechner, D., Siorpaes, K.: Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements. In: Workshop Semantic Wikis (2006)
8. Kavalec, M., Svátek, V.: A study on automated relation labelling in ontology learning. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.) Ontology learning from text: methods, evaluation and applications. IOS Press, Amsterdam (2005)
9. de Marneffe, C.-M., MacCartney, B., Manning, C.D.: Generating Typed Dependency Parses from Phrase Structure Parses. In: Proc. LREC 2006, pp. 449–454 (2006)
10. Milne, D.N., Witten, I.H.: Learning to link with Wikipedia. In: CIKM, pp. 509–518 (2008)
11. Pantel, P., Pennacchiotti, M.: Automatically Harvesting and Ontologizing Semantic Relations. In: [2]
12. Sánchez, D., Moreno, A.: Learning non-taxonomic relationships from web documents for domain ontology construction. Data Knowl. Eng. 64(3), 600–623 (2008)
13. Suchanek, F.M., Kasneci, M., Weikum, G.: Yago: a Core of Semantic Knowledge. In: Proc. WWW 2007, pp. 697–706 (2007)
14. W3C: Defining N-ary Relations on the Semantic Web. W3C Working Group Note (2006), http://www.w3.org/TR/2006/NOTE-swbp-n-aryRelations-20060412/
15. W3C: OWL Web Ontology Language. W3C Recommendation (2004), http://www.w3.org/TR/2004/REC-owl-features-20040210/