RC22868 (W0308-037) August 11, 2003
Computer Science

IBM Research Report

Towards Ontologies On Demand

Youngja Park, Roy J. Byrd, Branimir K. Boguraev
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Towards Ontologies On Demand

Youngja Park, Roy J. Byrd and Branimir K. Boguraev
IBM T.J. Watson Research Center
19 Skyline Dr., Hawthorne, NY 10562
{*young_park*}*@us.ibm.com*

Abstract

The Semantic Web aims at adding semantic knowledge into the web of natural language hypertext, enabling deep-level information search and information integration. However, building a knowledge base and an ontology is so costly and time-consuming that it hampers the progress of the Semantic Web activity.

We present a method for building *ontologies on demand* from scientific queries by applying text mining technologies. The method induces ontological concepts and relationships relevant to the query by analyzing search result documents together with domain-specific knowledge sources available on the Web. Users can use this partial ontology not only for ad-hoc search refinement but also for extending an existing domain ontology. The presented method can be used to produce, over several sessions, a personalized ontology.

Introduction

With the advance of the Internet and computer technology, we are living in an "information overloaded" world; thus, timely access to and digestion of information is increasingly challenging. Because on-line information is still only human-readable, humans have to manually browse through documents to find the information they want. Scientists, especially, find it difficult to stay up-to-date with the vast amount of new literature in their fields.

Recent research in the field of information retrieval (IR) aims to improve the keyword-based search by adding hypernyms or synonyms (Kerschberg, Kim, & Scime 2001; Mihalcea & Moldovan 2000; Voorhees 1994) from WordNet (Miller 1990) or frequently cooccurrent words (AltaVista 2002; Cooper & Byrd 1997) into the original query terms. These query expansion mechanisms generally improve the performance but suffer from ambiguities and inability to extend the underlying lexical resource.

Users need smarter and more versatile search capabilities through Semantic Web-enabled or ontology-enhanced search engines (Buttler *et al.* 2002; Guha, McCool, & Miller 2003; McGuinness 1999), which provide deeper understanding of domain concepts and their inter-relationships (Sheth, Arpinar, & Kashyap 2003). The Semantic Web is an extension of the current web in which information is given well-defined meaning, enabling computers and people to work in cooperation (Berners-Lee, Hendler, & Lassila 2001). The Semantic Web activity aims to add onto the existing web of hypertext a machine-readable semantic layer including the creation of semantic annotations and the linking of web pages to ontologies.

However, semantics-enabled intelligent search faces several challenging questions.

First, where are the ontologies and knowledge bases? A few human-made ontologies such as WordNet and Cyc (Lenat *et al.* 1990) exist. However, these general-purpose ontologies contain few scientific concepts—rendering them less useful for most scientific searches. The Metathesaurus in the UMLS[1] (Unified Medical Language System) contains more than 620,000 medical concepts (Humphreys, Lindberg, & Schoolman 1998). However, no reliable ontologies or knowledge bases exist for other fields. Furthermore, manual ontology construction demands a lot of time and effort from domain experts and ontology engineers. The WordNet, Cyc and UMLS projects consumed many person-years to come into existence.

Second, how can we keep an existing ontology, if available, up-to-date? New concepts and new instances and attributes of existing concepts are constantly introduced. For instance, an existing medical ontology might not contain SARS (Severe Acute Respiratory Syndrome), not to mention its relation with Hong Kong or Beijing. Wireless internet access was not a feature of a cellular phone a couple of years ago, but it is an important feature nowadays. In order to overcome the "knowledge-acquisition bottleneck", we need automatic or semi-automatic tools to build ontologies.

Recently, there have been efforts for building or maintaining ontologies semi-automatically from domain documents (Hahn & Schnattinger 1998; Kietz, Maedche, & Volz 2000). Advances in text mining technology have improved the automatic ontology construction process. However, the technology for automatic ontology construction is still in its infancy due to the problems of

[1] http://www.nlm.nih.gov/research/umls

deep-level human language understanding. Much further research is needed to achieve the goal of true automatic ontology construction. Difficulties, to name a few, are:

- recognizing domain concepts that are worthy of inclusion in an ontology,
- defining a set of relationships for the concepts, and
- identifying the relationships in natural language text.

Semantics-enabled search requires ontologies, but neither the automatic nor the manual approach to *full* ontology construction looks promising in the near future.

In this paper, we propose *ontologies on demand* as a way to close this gap. This method builds an ontology from scientific queries by applying text mining technologies. The system processes documents returned by a search engine to find terms semantically related to the target query. It also identifies the relationships in which they participate. We call the ontology a *partial* ontology because it defines only the concepts represented by the query terms.

We argue that this approach is more feasible than trying to build a *full* ontology from a collection of documents for the following reasons. First, the system intends to focus on a small number of domain concepts that are denoted by the query terms. Thus, identifying target concepts and relations in documents is easier. Second, the search result contains documents regarding specific concepts. We expect much less semantic ambiguity in the documents, which makes the approach more practical.

The resulting ontology describes the target terms in a semantic space, and provides users with a deeper understanding of the query. The users can use the ontology not only for performing ad-hoc search refinement but also for building or extending a domain ontology. Scientists' searches tend to focus on new concepts or important issues in the field; thus, the up-to-date ontologies provided by *ontologies on demand* will be very valuable in the scientific domain.

Scenario and System Design

This system utilizes scientific search activities for creating a personalized domain-specific ontology. In the Introduction, we contrasted our goal of building *ontologies on demand* with automatic contruction of full ontologies, pointing out that the former would be more feasible. In addition, we are addressing the task of building ontologies in the presence of the search user (a domain expert) or an ontology engineer. In this environment, the system can tolerate less precise extraction methods, because a human user can interactively control the final decision.

Figure 1 shows the overall architecture of the system. A search engine extracts documents that match the user's search terms from the given domain corpus or the Web. To augment the ontology builder's input, the system can add domain glossaries, which are located through a glossary search on the Internet[2], to the top n ranked documents. The ontology builder, then, recognizes terms semantically related to the target terms and ontological relations involving them. That is, we use the query terms as anchors for the target concept and extend the concept space with other semantically-related terms.

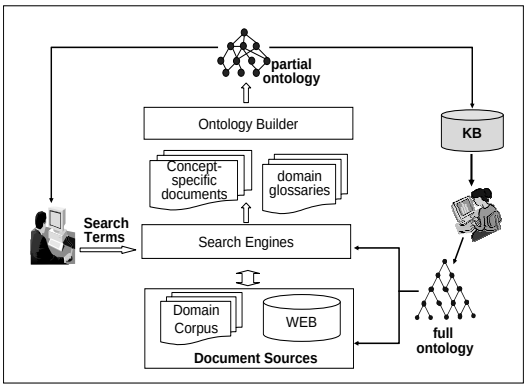

Figure 1: System overview for *ontologies on demand*

Research query terms and the returned documents are good knowledge sources for domain concepts. However, not even authors of domain-specific documents explicitly define *all* their terms, because the concepts are well-known or they can be inferred from the context. Therefore, the method described in this paper consults external domain knowledge sources, such as glossaries, in addition to the search results, to find definitional knowledge and ontological relations for the target concept.

The system allows users to exploit the ontologies in two modes. First, they can use the ontology for ad-hoc search refinement. Through the ontology, users will have deeper understanding of the query terms and can iteratively modify the query to better represent their information need. The system generates a new ontology from the new search and extends the previous one if the two searches are related.

Second, after the search, the ontologies are saved into a knowledge base, and an ontology engineer can incrementally build a domain ontology. Repeating this process will produce a personalized ontology for the scientist or the research group.

Ontology Construction

Our on-going work addresses two major challenges to achieve search-directed ontology construction. First, we recognize domain-specific concepts that are relevant to the user search. Recognizing all nouns and verbs in the

[2]http://labs.google.com/glossary

documents is insufficient because domain-specific documents contain many generic terms as well as conceptually relevant terms.

Second, we discover ontological *named* relationships between domain concepts from unstructured text. Previous approaches concentrate on identifying taxonomic relations (e.g., IS-A relation) by using statistical or linguistic information (Hahn & Schnattinger 1998; Hearst 1992; Pereira, Tishby, & Lee 1993). Maedche and Staab (Maedche & Staab 2000) presents a way to find non-taxonomic conceptual relations from text but the relationships remain unnamed.

Search-Directed Entity Recognition

The recognition of terms conceptually relevant to the query begins with the identification of glossary items in text (e.g., the search documents and domain glossaries in this work). Glossary items are words or phrases which describe the domain concepts. The identified glossary items are evaluated to select items semantically relevant to the query terms. In the relation extraction step, only these selected glossary items are considered as the target concepts.

Our glossary extraction system comprises term recognition, modifier filtering, glossary item aggregation and confidence computation (see (Park, Byrd, & Boguraev 2002) for complete descriptions).

Term recognition step identifies single- and multi-word phrases (noun phrases) by applying a FST-based noun phrase recognizer. In addition, we process out-of-vocabulary words to recognize more domain-specific technical words (Park 2002). Note that many technical words are missing from dictionaries.

Modifier filtering distinguishes domain-specific modifiers from generic modifiers. Many domain-specific noun phrases contain generic modifiers, which do not contribute to the domain concept. For instance, "psychiatric" in "psychiatric disorder" is domain-specific, but "related" in " related disorder" is not domain-specific. The system filters out generic modifiers based on statistical information.

Glossary item aggregation combines variations of expressions into a single item. Technical documents typically contain variations such as abbreviations ("International Dyslexia Association" and "IDA"); spelling errors or alternative spellings ("anesthesia" and "anaesthesia"); and orthographic variants ("Attention-Deficit Hyperactivity Disorder", "attention deficit hyperactivity disorder" and "attention deficit/hyperactivity disorder"). We recognize abbreviations and their definition in documents by applying *abbreviation pattern-based rules* as well as textual cues or keywords (Park & Byrd 2001). Spelling errors or alternative spellings are determined based on string edit distance of two words.

Finally, confidence values for the items are computed based on the domain-specificity and term cohesion (for multi-word terms). Domain-specificity denotes how strongly a term is related to the domain and is computed by the relative probability of occurrences in domain text and in general text. Term cohesion represents the tendency of words in a multi-word term appearing together in the term. If the confidence value of a term is higher than a given threshold, it is considered domain-specific.

We evaluated glossaries extracted from document collections using these methods. The methods were especially successful when extracting glossaries from technical literature. In that work, we suggested that glossary items so derived are suitable for use as names of domain-specific concepts. In our present ontology work, we explore and exploit that suggestion by mapping glossary items to query terms and to ontology concepts.

In a further step, we aim to select glossary items semantically related to the search terms from among the recognized items. We regard terms t_1 and t_2 to be semantically related if

- t_1 is found in the glossary definitions for t_2,
- t_1 and t_2 appear together in certain syntactic structures (e.g., apposition and conjunction), or
- t_1 appears many times in documents resulting from a search for t_2

The following relation extraction procedure only concentrates on these relevant terms recognized by this step.

Relation Extraction

The relation extraction module extracts "IS-A", "Alias", and other named relations from syntactic dependency relations involving the query terms and the glossary items conceptually relevant to the query. Syntactic dependency relations coincide closely with semantic relations between the entities (Maedche & Staab 2000). We process documents with a syntactic parser (McCord 1990) to obtain grammatical dependency relations of constituents and to recognize patterns for relations.

IS-A. IS-A, or hyperym/hyponym, relations are extracted in two ways. First, we search glossary definitions for the query term and process the definitions to find the genus term. The genus term in the first sentence of a gloss usually represents the hypernym of the term defined (Vossen, Meijs, & den Broeder 1989). For instance, we conclude *dyslexia* is a *language-based learning disability* from the glossary definition.[3]

The common syntactic structures in which the genus terms are found are:

- $noun_0$ which ....
- a {kind|type|category} of $noun_0$ ...
- a {term|concept} { [used] to verb|for verb-ing} ...

[3]Dyslexia is a language-based learning disability in which a person has trouble understanding words, sentences, or paragraphs.

Second, we recognize lexico-syntactic patterns in search documents and other parts of glossary definitions, which indicate hyponym relations or definitional sentences. Example patterns are shown in Table 1. When these and similar syntactic patterns

$noun_1$ is {a|the} $noun_0$
$noun_1$ is a term {[used] to verb|for verb-ing} $noun_0$
such $noun_0$ as $noun_1$, $noun_2$,..., {and|or} $noun_n$
$noun_0$ {including|especially} $noun_1$,...,{and|or} $noun_n$
$noun_1$, $noun_2$,..., $noun_n$,..., {and|or} other $noun_0$
$noun_0$ except $noun_1$, $noun_2$,..., {and|or} $noun_n$
$noun_0$, for example $noun_1$, $noun_2$,..., {and|or} $noun_n$

Table 1: The lexico-syntactic patterns for hypernym/hyponym relation (motivated by the work in (Hearst 1992)). The patterns indicate that $noun_i, 1 \leq i \leq n$, are hyponyms of $noun_0$.

occur in text, we typically find that the $noun_i, 1 <= i <= n$, are hyponyms of $noun_0$. In these patterns, recall that $noun_i$ actually refers to both single- and multi-word nominal expressions and at least one of $noun_i$s belongs to the selected glossary items.

Alias. This relation specifies alternative names for a concept. Abbreviations are the most common examples of this relation. The system for matching abbreviations and their definitions, which was described in the previous section, is used for this purpose (Park & Byrd 2001).

We also identify patterns for recognizing other aliases as shown in the following examples.

- *Zomig*, formerly known as *311C90*
- *3,4-methylenedioxymethamphetamine* (also known as "*Ecstasy*")

Verbal Relations. Syntactic structures are distinct from semantic structures. Nevertheless, in general, lexical items which express predicate relations take, as their syntactic dependents, nominal expressions which name the predicate's arguments. These predicate-argument structures can often be interpreted as expressing relationships. A common example is that the subject and object of a verb are the participants in the relation expressed by the verb. We find the dependency relations in which the target terms or the selected glossary items appear and generate relations named with the verbs.

Table 2 and Table 3 show examples of the selected glossary items and some relations respectively. These examples are extracted from glossary definitions and search results from MEDLINE abstracts database for *dyslexia*.

Future Research for Ontology Extension

The preceding discussion illustrates how search and question-answering systems based on ontologies offer opportunities for system users to build *partial* (or "session-specific") ontologies for special purposes and

alzheimer	auditory processing disorder
children	developmental delay
disability	developmental dyslexia
disease	language impairment
sclerosis	speech milestone

Table 2: Glossary items semantically related to *dyslexia*. The glossary items are extracted the search results (17 MEDLINE abstracts) for *dyslexia*

to correct and extend an underlying *full* ontology. In this section, we give three scenarios for using interactive tools to support this activity and we describe our plans for building those tools.

Class Assignment

In the scenario for this tool, a user has encountered (perhaps in a document) or entered (perhaps in a query) a lexical item which is unknown to the system. This means that the lexical item does not name any known concept in the ontology. The objective of *class assignment* is to assign the item to a known ontology class and to characterize that decision with a measure of the system's confidence in its assignment. Under user control, the ultimate goal is to modify the (*partial* and/or *full*) ontology by entering the item as a new member of its assigned class.

Several technologies cooperate to provide the function described. In one, a machine-learning classifier for text mentions (Ando, personal communication), once trained with a "gold standard" corpus containing annotations for a class of interest, examines lexical items in unlabeled text and labels some of them as mentions of the class.

If the classifier fails to classify the unknown item, either because of low confidence or absence of a suitable classifier, other methods are applied. These methods depend on a set of text analysis techniques for finding hypernym relations between items in text. In one method, the unknown word is used as a query against a suitable document collection, such as the Web or a domain-specific corpus. The documents in the query result are grouped by using a clustering algorithm (Ando 2000), in an attempt to tease out any ambiguity of the original item. Syntactic patterns designed to identify hypernym and other relations are applied to the documents in each cluster, as described in the previous section. Then, identified relations linking the unknown item to items mentioning known ontological classes are presented to the user.

Class Creation

The system user may also want to define a new class of entities sharing some properties and to add the new class to the ontology. To do so, the user begins by giving the system a small set of lexical items naming sample class members. The system uses the sample class members to produce suggestions of "more items like this" for

$entity_1$	relation	$entity_2$
dyslexia	IS-A	learning disability
dyslexia	IS-A	reading disorder
developmental dyslexia	IS-A	impairment of reading skills
DAAT	IS-A	treatment
Attention Deficit-Hyperactivity Disorder	HasAlias	ADHD
cortical multiple sclerosis	CAUSE	handicap
dyslexics	SHOW	auditory and visual abnormality
mental disorder	CORRELATE_WITH	CAPD
isoprenoid pathway	PRODUCE	metabolite
isoprenoid pathway	PRODUCE	digoxin
isoprenoid pathway	PRODUCE	dolichol
isoprenoid pathway	PRODUCE	ubiquinone

Table 3: Sample relations extracted from a search for *dyslexia*. These relations are extracted from glossary definitions for *dyslexia* and search results from MEDLINE abstracts

the user to validate. After validation, the new class and its members are inserted into the ontology. As optional steps, the new class may be related to existing classes and a classifier capable of recognizing text mentions of further members of the new class may be created.

The techniques for realizing this *class creation* scenario begin with a set of standard text analysis tools (Neff, Byrd, & Boguraev 2003; Park, Byrd, & Boguraev 2002) for identifying possible lexical items and their syntactic contexts in a large corpus of domain texts. The items are represented as vectors of features which characterize the contexts of the items' occurrences in the corpus. One of two methods are used to grow the set of seed items. In the first (Ando, personal communication), vector-space methods find additional lexical items which are close to the seeds and presents them to the user in the order of closeness. In the second method, inspired by (Thelen & Riloff 2002), additional items which are most similar to the current set of class members are presented to the user in an incremental bootstrapping procedure.

Relation Discovery

In this scenario, the user wishes to add to the ontology binary relations that can link members of two given classes. The system responds by inducing a set of relations from relevant documents and organizing them for presentation to the user.

Relation Discovery begins by selecting a set of document sentences in which lexical items from extensions of the two selected classes occur together. It then uses relation extraction methods, such as those described in the previous section, to extract candidate relations having the two items as arguments. Since our extraction algorithms may extract many similar relations, we organize the candidates by clustering them with respect to their arguments, contexts, and extraction patterns. The system may optionally name the clusters or prompt the user to do so.

To further organize and describe the candidate relations, the system finds metarelations such as "same-as", "special-case-of" and "inverse-of" by examining the sets of relation instances found by relation extraction and the recognition grammars. In heuristics used to do this, the system inspects the set of tuples (pairs, for binary relations) containing the fillers of all relation instances for two test relations. If the sets of tuples have a large overlap, then the two relations might be the "same-as" one another. If one set of tuples largely includes the other set, then the second relation is suggested to be a "special-case-of" the first. Finally, if the inverse of the first set of (ordered) tuples has a large overlap with the second set, then the two relations might be "inverses-of" one another.

Summary

In this paper, we argue for the utility — both short-term and strategic — of narrowly focused ontological structures, capable of acting as a prompting device in the process of query construction, as well as an organic extension of broader and more persistent (scientific) ontologies

We presented an automatic tool to build partial ontologies on-the-fly from the results of users' search queries and domain glossaries. The ontologies give the users a deeper understanding of the query terms and help them in refining their searches. In addition, a domain expert can build a domain ontology or extend an existing ontology with these partial ontologies. Note that frequent search terms reflect current interests in the domain; therefore, this system helps to keep an ontology up-to-date.

Most of the technologies described, or referred to, earlier in this paper as crucial to implementing *ontologies on demand* have been independently evaluated (see (Neff, Byrd, & Boguraev 2003) and citations therein for more details on individual performance figures). A remaining challenge, however, is to design and carry out evaluations of the heuristic methods that combine them. It is always possible to adopt a task-based approach to evaluation of the improvements our methods produce in search users' productivity will be possible. It will, however, be expensive. Thus, we will seek more automated methods for evaluating subcom-

ponents of our overall system that do not require large investments of human time and effort.

References

AltaVista. 2002. Altavista prisma. `http://www.altavista.com/prisma`.

Ando, R. K. 2000. Latent semantic space: Iterative scaling improves inter-document similarity measurement. In *Proceedings of SIGIR'2000*, 216–223.

Berners-Lee, T.; Hendler, J.; and Lassila, O. 2001. The semantic web. Scientific American.

Buttler, D.; Coleman, M.; Critchlow, T.; Fileto, R.; Han, W.; Liu, L.; Pu, C.; Rocco, D.; and Xiong, L. 2002. Querying multiple bioinformatics information sources: Can semantic web research help? SIGMOD Report.

Cooper, J., and Byrd, R. J. 1997. Lexical navigation: visually prompted query expansion and refinement. In *Proceedings of the second ACM international conference on Digital libraries*, 237–246.

Guha, R.; McCool, R.; and Miller, E. 2003. Semantic search. In *Proceedings of WWW2003*.

Hahn, U., and Schnattinger, K. 1998. Ontology engineering via text understanding. In *Proceedings of the 15th World Computer Congress*.

Hearst, M. 1992. Automatic acquisition of hypernyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*.

Humphreys, B.; Lindberg, D.; and Schoolman, H. 1998. The unified medical language system: An informatics research collaboration. *Journal of the American Medical Informatics Association* 5(1):1–13.

Kerschberg, L.; Kim, W.; and Scime, A. 2001. A semantic taxonomy-based personalizable meta-search agent. In *Proceedings of the 2nd International Conference on Web Information Systems Engineering (WISE)*.

Kietz, J.-U.; Maedche, A.; and Volz, R. 2000. A method for semi-automatic ontology acquisition from a corporate intranet. In *Workshop "Ontologies and Text", co-located with EKAW2000*.

Lenat, D.; Guha, R. V.; Pittman, K.; Pratt, D.; and Shepherd, M. 1990. Cyc: Toward programs with common sense. *Communications of the ACM* 33(8):30–49.

Maedche, A., and Staab, S. 2000. Discovering conceptual relations from text. In *Proceedings of ECAI*.

McCord, M. 1990. Slot grammar: A system for simpler construction of practical natural language grammars. In Studer, R., ed., *Natural Language and Logic: International Scientific Symposium*, 118–145. Springer Verlag.

McGuinness, D. L. 1999. Ontology-enhanced search for primary care medical literature. In *Proceedings of the International Medical Informatics Association Working Group 6- Medical Concept Representation and Natural Language Processing Conference*.

Mihalcea, R., and Moldovan, D. 2000. Semantic indexing using wordnet senses. In *Proceedings of ACL Workshop on IR & NLP*.

Miller, G. 1990. Wordnet: an on-line lexical database. *International Journal of Lexicography* 3(4).

Neff, M. S.; Byrd, R. J.; and Boguraev, B. K. 2003. The Talent system: Textract architecture and data model. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)*.

Park, Y., and Byrd, R. J. 2001. Hybrid text mining for matching abbreviations and their definitions. In *Proceedings of Empirical Methods in Natural Language Processing*, 126–133.

Park, Y.; Byrd, R. J.; and Boguraev, B. K. 2002. Automatic glossary extraction: Beyond terminology identification. In *Proceedings of the Nineteenth International Conference on Computational Linguistics*.

Park, Y. 2002. Identification of probable real words: An entropy-based approach. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, 1–8.

Pereira, F. C. N.; Tishby, N.; and Lee, L. 1993. Distributional clustering of english words. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 183–190.

Sheth, A.; Arpinar, I. B.; and Kashyap, V. 2003. Relationships at the heart of semantic web: Modeling, discovering, and exploiting complex semantic relationships. In Nikravesh, M.; Azvin, B.; Yager, R.; and Zadeh, L. A., eds., *Enhancing the Power of the Internet: Studies in Fuzziness and Soft Computing*. Springer Verlag.

Thelen, M., and Riloff, E. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

Voorhees, E. M. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. (Special Issue of the SIGIR Forum)*, 61–69.

Vossen, P.; Meijs, W.; and den Broeder, M. 1989. Meaning and structure in dictionary definitions. In Boguraev, B., and Briscoe, T., eds., *Computational Lexicography for Natural Language Processing*, 171–192. Longman.