Department of
Information Engineering
and Computer Science DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

SUMMARIZATION OF CONCEPTS FOR VISUAL DISAMBIGUATION

Aliaksandr Autayeu, Pierre Andrews and Fausto Giuchiglia

April 2011

Technical Report # DISI-11-461

Summarization of Concepts for Visual Disambiguation

Aliaksandr Autayeu Pierre Andrews
Fausto Giunchiglia
University of Trento, Italy
{autayeu;andrews;fausto}@disi.unitn.it

Abstract

Controlled vocabularies that power semantic applications allow them to operate with high precision, which comes with a price of having to disambiguate between senses of terms. Fully automatic disambiguation is a largely unsolved problem and semi-automatic approaches are preferred. These approaches involve users to do the disambiguation and require an adequate user interface. However, term definitions are usually lengthy and not only occupy valuable screen space, but reading and understanding these definitions requires the user's attention and time. As an alternative to using definitions we propose to use a summary — a "single word" disambiguation label for a concept.

In this paper we present an algorithm to summarize concepts from a controlled vocabulary. We evaluate the algorithm with 51 users and show that the algorithm generates summaries that have good discriminative and associative qualities. In addition, the length of summaries are comparable to the length of the original terms, thus making the algorithm particularly useful in situations where screen estate is limited.

1 Introduction

Semantic applications need natural language words to have precise meanings. These meanings are based on a controlled vocabulary that serves as a reference for word senses. However, using a controlled vocabulary where words might have multiple senses requires to choose between those senses. This task is well known in Natural Language Processing as Word Sense Disambiguation (WSD).

Fully automatic Word Sense Disambiguation is considered in the field to be particularly hard, as was already pointed out in the state of the art [5]. It is indeed difficult to reach precision values over 66%. However, we are interested in *semi-automatic* methods that can help the users in providing semantic annotations. Thus, a good way to approach the WSD issue is to provide an adequate user interface so that the users can disambiguate the right sense for a term directly when providing some inputs (such as tag annotation of images). However, when dealing with polysemous words, displaying a simple term is not enough and a visual disambiguation of the term is required.

Currently, the only available way to perform such disambiguation is by displaying the definition of the term. For instance, for the “java” term, WordNet [3] provides the senses:

- Java (an island in Indonesia to the south of Borneo; one of the world’s most densely populated regions)
- coffee, java (a beverage consisting of an infusion of ground coffee beans) ”he ordered a cup of coffee”
- Java (a platform-independent object-oriented programming language)

Displaying such long definitions in a user interface is not always practical as the screen real estate is limited. In this paper, we propose a *summarization* algorithm that can produce a “single word” disambiguation label for a term. For instance, for the “java” term, we would have the three following summaries:

- Java – island
- java – beverage
- Java – programming

These disambiguating labels do not provide a full definition, but are sufficient for the user to disambiguate between all the ambiguous senses of a single word.

2 Algorithm

We consider that each Part Of Speech (POS) – such as noun, verb, adjective, adverb – can have a different summary, and that the summarization algorithm might need different heuristic depending on the POS. We thus explore the summarization algorithm considering the particular characteristics of each of them. In the following paragraphs we provide the characteristics of each of the POS, and report possible heuristics based on WordNet 2.1.

2.1 Noun summarization

In WordNet there is a total of 117 097 nouns, out of which 13.47% have more than one sense, referring to 43 783 different senses, giving us an average of 2.7 senses per ambiguous noun. To create a summary label for a given ambiguous noun, we defined four heuristics, choosing (in the presented order) the first available heuristic among the following:

1. return the first shortest unused lemma among the available words of the same synset;
2. return the first shortest unused lemma among the available words in the hypernym’s (parent) synset;

3. return the first shortest unused lemma among the available words in the hyponyms' (child) synsets;

4. if there are no hypernym synsets available, return the noun itself (original token)

We call a "used lemma" a sequence of characters, which already serves as a summary for another sense. We always use the "first *un*used lemma" when possible as different senses might have the same terms in their synset and thus we choose only the first term we can find that has not yet been selected as a summary for another of the senses.

We check the length of the lemma and choose the shortest among the available choices for two main reasons: first, very often the shortest word is the simplest one and second, to save screen space (changing as little as possible the normal flow of the annotation process). There are a few cases (1.48%) where several senses share the same summary; in particular, some of them (0.82%) have the same summary for all the senses – in these cases it is impossible to help the user disambiguate with such a summary and the definition of the synset will have to be used. Roughly half of the ambiguous noun senses produce a shorter summary label, on average 2.31 characters, and in the other half of the cases (22 001 out of 43 783 senses) the summary label produced is longer than the original word, namely, on average 4.84 characters longer. Table 1 includes some examples of summary labels for nouns.

2.2 Verb summarization

In WordNet there is a total of 11 488 verbs, out of which 45.49% have more than one sense, referring to 18 629 different senses, giving us an average of 3.56 senses per ambiguous verb. To create a summary label for a given ambiguous verb we defined four heuristics, choosing (in the presented order) the first available heuristic among the following:

1. return the first shortest unused lemma among the available words of the same synset;

2. return the first shortest unused lemma among the available words in the hypernym's synset;

3. return the first shortest unused lemma among the available words in the hyponyms' synsets;

4. return the first word of the gloss (often a well-known verb such as "to cause", "to have" or "to be").

In a few cases (1.38%), several senses share the same summary while some terms (0.23%) have the same summary label for all senses. In a majority of cases (65%) the summary label is, on average, 1.84 characters shorter, while in 35% of cases the produced summary label is, on average, 2.59 characters longer. Table 1 includes some examples of summary labels for verbs.

Sense	Gloss	Obtained via	Summary
triangle#1	a three-sided polygon	SYNSET	trigon
triangle#2	something approximating the shape of a triangle;	HYPERNYMS	form
triangle#3	a small northern constellation near Perseus between Andromeda and Aries	SYNSET	Triangulum
add#1	make an addition (to); join or combine or unite with others; increase the quality, quantity, size or scope of; “We added two students to that dorm room”	HYPERNYMS	increase
add#2	state or say further; “‘It doesn’t matter,’ he supplied”	SYNSET	append
add#3	bestow a quality on; “Her presence lends a certain cachet to the company”	SYNSET	lend
aboriginal#1	of or pertaining to members of the indigenous people of Australia; “an Aboriginal rite”	PERTAINYM	Abo
aboriginal#2	characteristic of or relating to people inhabiting a region from the beginning; “native Americans”; “the aboriginal peoples of Australia”	SYNSET	native
aboriginal#3	having existed from the beginning; in an earliest or original stage or state; “aboriginal forests”; “primal eras before the appearance of life on earth”; “the forest primeval”	SYNSET	primal
aboard #1	part of a group; “Bill’s been aboard for three years now”	GLOSS	part
aboard #2	on a ship, train, plane or other vehicle	SYNSET	onboard
aboard #3	on first or second or third base; “Their second homer with Bob Allison aboard”	SYNSET	on_base
aboard #4	side by side; “anchored close aboard another ship”	SYNSET	alongside

Table 1: Summarization Examples for Nouns, Verbs, Adjectives, Adverbs

2.3 Adjective summarization

In WordNet there is a total of 22 141 adjectives, out of which 23.72% have more than one sense, referring to 14 413 different senses, giving us an average of 2.7 senses per ambiguous adjective. To create a summary label for a given ambiguous adjective we defined six heuristics, choosing the first available heuristic among the following:

1. return the first shortest unused lemma among the available words of the same synset;
2. return the first shortest unused lemma among available words in the satellite synsets (using the `similar_to` relation);
3. return the first shortest unused lemma among the available words in the pertainyms' (related noun) synsets;
4. return the first shortest unused lemma among the available words in the `see_also` synsets;
5. return the first shortest unused lemma among the available words in the antonyms' synsets;
6. return the first word of the gloss.

There are few cases (0.36%) where several senses share the same summary label, some of them (0.11%) have the same summary label for all senses. In a majority of cases (56%) the summary label is on average 2.05 characters shorter than the original word. In 44% of the cases the summary label is on average 2.89 characters longer than the original word. Table 1 includes some examples of summary labels for adjectives.

2.4 Adverb summarization

In WordNet there is a total of 4 601 adverbs, out of which 16.32% have more than one sense, referring to 1 870 different senses, giving us an average of 2.49 senses per ambiguous adverb. To create a summary label for a given ambiguous adverb we defined four heuristics, choosing the first available heuristic among the following:

1. return the first shortest unused lemma among the available words of the same synset;
2. return the first shortest unused lemma among the available words in derived synsets of this adverb (using the `derived_from` relation);
3. return the first shortest unused lemma among the available words in the antonyms' synsets;
4. return the first word of the gloss.

	total	with multiple senses, %	number of senses	ambiguity	overlapping summaries, %	no distinct summaries, %
Nouns	117 097	13.47	43 783	2.70	1.48	0.82
Verbs	11 488	45.49	18 629	3.56	1.38	0.23
Adjectives	22 141	23.72	14 413	2.70	0.36	0.23
Adverbs	4 601	16.32	1 870	2.49	6.66	2.40

Table 2: WordNet Word Ambiguity per POS

There are a more cases than for the other POS (6.66%) where several senses share the same summary and some of the adverbs (2.4%) have the same summary label for all the senses. This is due to the lower number of relations for the adverbs in WordNet. In a majority of cases (66%) the produced summary label is on average 2.56 characters shorter than the original word. In 34% of the cases the summary label is on average 3.24 characters longer than the original. Table 1 includes some examples of summary labels for adverbs.

3 Evaluation

3.1 Scenario

As ultimately these summaries should be used to disambiguate terms visually when the user is providing a controlled annotation, we evaluated their quality and precision by evaluating them directly with the users. For each word, we evaluated all the defined heuristics applicable to the POS this word belongs to. This gave us a comparable measure of the quality of each heuristic for each POS.

The evaluation methodology is based on the fact that the purpose of the summary labels is twofold:

1. the summary should represent the meaning of the gloss,

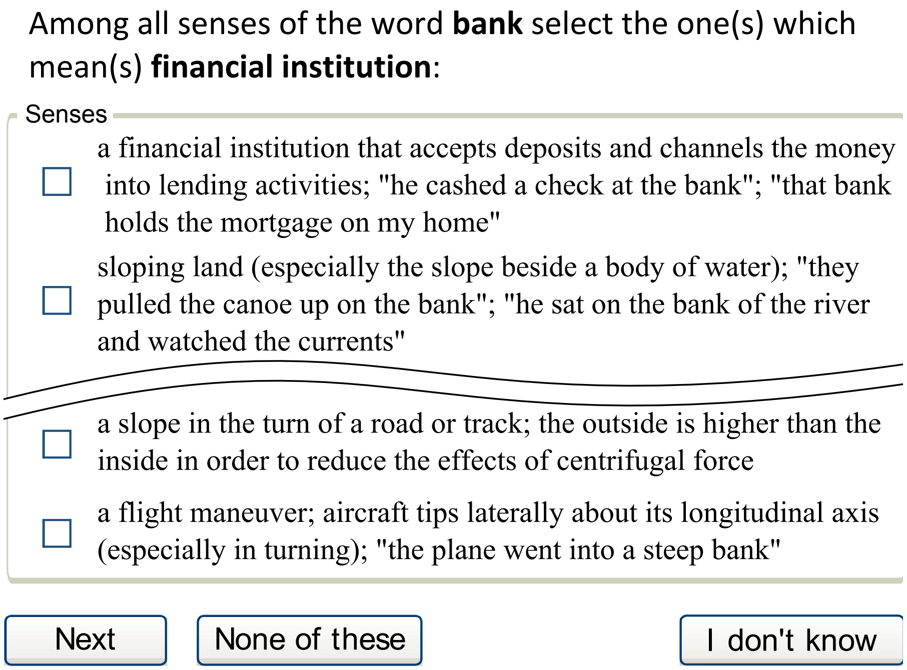
2. it should discriminate well enough between the difference senses of the same word.

We have thus introduced two scenarios to evaluate the summary heuristics.

Scenario 1 The user is presented with a word and its summary in the form of the following question:

> Among all these senses of the word **word** select the one(s) which mean(s) **summary**.

The user is then presented with a list of senses that are expressed by their glosses obtained from WordNet. The user is asked to select the right sense(s) corresponding to the summary shown. In case of doubts on the answer, the user also has the option of skipping the current question.



Among all senses of the word **bank** select the one(s) which mean(s) **financial institution**:

Senses

□ a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home"

□ sloping land (especially the slope beside a body of water); "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"

□ a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force

□ a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); "the plane went into a steep bank"

Figure 1: Example of the first validation scenario question for the word "bank"

The example of the first validation scenario provided in Figure 1 illustrates the rationale behind allowing the user to choose multiple senses as an answer to a question. In WordNet, there are senses that may be too fine-grained and the user may perceive the produced summary label as related to several displayed definitions. These too fine grained senses in WordNet might decrease the disambiguation precision of the summary label during the annotation task.

To assess the quality of the results of the first validation scenario we have devised six answer categories. These answer categories are not mutually exclusive and serve to demonstrate different properties of the summary heuristic in question. The categories and their purpose are:

unknown when the user clicked "I don't know" button. This category shows that the user is not familiar with the presented word or summary label;

none when the user clicked "None of these" button. This category shows that user was not able to associate the summary with a sense;

correct when the user selected one sense and this sense is correct. This category shows that the summary label is good, as the user was able to associate it with the correct sense;

semicorrect when the user selected more than one sense, and the correct one is among them. This category shows that the summary is potentially good, because the user was able to associate it the correct sense, however, the user was not able to make a proper distinction;

incorrect when the user selected only incorrect senses. This category shows that the summary label is potentially bad, because the user was not able to associate the correct sense; and

more than 1 selected sense when the user selected more than one sense. This category shows the cases where the senses are too fine-grained and confuse the user.

Scenario 2 We test the *discrimination power* of summary labels to disambiguate between different senses without displaying a definition. To ease the cognitive load on the users, and thus simplify the validation task, we do this for pairs of senses of a word, instead of all senses at once, by asking the following question:

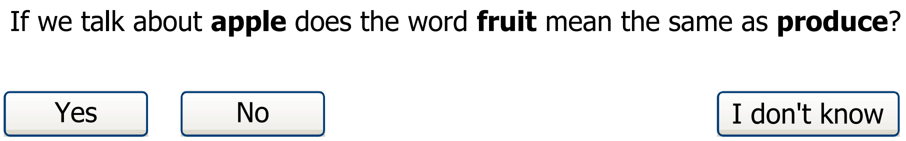
> If we talk about **word**, does the word **summary#sense1** mean the same as **summary#sense2**?

The user is then given the choice between three answers (as illustrated in Figure 2):

Yes means that the user understood the question but the discrimination is bad (i.e., the two summaries mean the same thing to the user);

No means that the user understood the question and the discrimination is good (i.e., the two summaries represent different senses); and

I don't know means that the user did not understand something in the question and could not answer the question.



If we talk about **apple** does the word **fruit** mean the same as **produce**?

Figure 2: Example of the second validation scenario's question for the word "apple" with summaries "fruit" and "produce"

Given that the answers to these questions can be subjective or error prone (depending on the summaries generated, some questions were quite ambiguous), we decided to evaluate the agreement between different users for the same exact question. We thus showed each instance of the questions to more than one user so that the agreement between the users could be computed (see Section 3.4) and therefore, we could guarantee a certain accuracy for the results.

3.2 Dataset

POS	Summary count	"Frequency" of use
ADJECTIVE	1 104	5.71
NOUN	4 318	26.86
ADVERB	271	16.07
VERB	3 621	11.00
TOTAL	9 314	

Table 3: Summaries by Part of Speech (POS)

The summarization algorithm was tested on a subset of WordNet (see Table 3 for details). Before conducting the evaluation we performed several dryrun evaluations with a limited set of users to test the phrasing of the questions and the design of the experiment.

In many cases, some users found the questions difficult to understand or to answer. The major reason for this difficulty was that, in some parts of WordNet, there is a very fine grained definition and specification of senses.

To tackle this issue, we have exploited the *use count* that is provided on a subset of WordNet and describes the "frequency of use" of the terms in a representative corpus of English. We have thus generated summaries only for words having non-zero use count in WordNet. This limited the questions to the most used words and thus, potentially, better known words, addressing the limited vocabulary problem to some extent.

Heuristic	Summary count	"Frequency" of use
CHILD	1 253	21.17
DERIVED	125	7.97
GLOSS	1 300	13.08
PARENT	3 804	24.57
SIMILAR_TO	756	6.58
SYNSET	2 076	11.24
TOTAL	9 314	

Table 4: Summaries by Heuristics

It is worth noting that different heuristics apply in different proportions to each POS (see Table 4 for details):

1. The "Synset" heuristic applies to all parts-of-speeches;
2. the "Parent" and the "Child" heuristics apply to both nouns and verbs;
3. the "Gloss" heuristic applies only to verbs due to the way in which the glosses are written;

4. the “Derived” heuristic applies only to adverbs; and
5. the “Similar_to” applies only to adjectives.

3.3 Participants

51 users participated in the evaluation, including representatives of both genders, various age groups (between 20 and 60) and various cultures. While the majority of the users were fluent non-native English speakers, some native speakers as well as bilingual users participated in the evaluation. Fifteen users answered more than 100 questions each, and a top contributor answered 700 questions. 25 users answered at least 40 questions each while eleven users answered less than 20 questions each. On average we collected 83 answers per user.

3.4 Users’ Agreement

To ensure the quality of the validations, a subset of the questions were handed out to at least two different users. This is a standard procedure in the construction of language datasets [2], as it allows the evaluations of the validity and reproducibility of the results in annotation tasks that might be ambiguous due to their natural language nature. It is accepted, in the state of the art, that presenting all the questions to every user would require too much resources and thus, a representative subset of the questions needs to be validated by more than one user.

We have collected 308 double-rated questions for the first scenario and 301 questions for the second scenario. To compute the user-agreement, we used the “agreement without chance correction” [2] where we considered an answer to a question as an item. One question could have more than one answer, each given by a different user. It is also useful to keep in mind that the questions from the first scenario can have multiple independent answers and as such are more difficult to agree upon for the users. Tables 5 and 6 provide details on the overall proportion of agreement by the type of answers across all heuristics.

Answer Type	Scenario 1	Scenario 2
UNKNOWN	0.18	0.23
NONE	0.23	n/a
CORRECT	0.59	0.72
SEMICORRECT	0.04	n/a
INCORRECT	0.35	0.43

Table 5: User agreement by question and answer type in the first and second scenarios.

3.5 Precision Results

The first scenario allows us to measure the precision of the summarization heuristics. That is: given its summary, with what precision can a user select the right definition

Answer type	child	derived	gloss	parent	similar_to	synset
UNKNOWN	0.50	0.50	0	0.13	0	0
NONE	0.50	0	0	0.09	0.40	0
CORRECT	0.37	0	0.53	0.62	0.61	0.66
SEMICORRECT	0	0.09	0.05	0	0	0.07
INCORRECT	0.50	0.50	0	0.30	0.36	0.40
UNKNOWN	0.18	1	0.25	0.25	0	0
CORRECT	0.82	0	0.72	0.75	0.60	0.53
INCORRECT	0	0	0.47	0.38	0.66	0.50

Table 6: User agreement by heuristic and answer type for questions in the first and second scenarios.

for a word within all its senses?

Table 7 provides a detailed view of the distribution of answers per POS, heuristic and type of answers given by the user. As mentioned earlier, each heuristic applies to a different proportion (if it applies) for each POS category and thus it is more interesting to consider the results separately. Note that the types of answers are not exclusive, in fact *more than one selected sense* includes some of the *incorrect* answers[1] and the *semicorrect* ones[2].

	Child		Derived	Gloss	Parent		Similar To	Synset			
Answer Type	N	V	R	V	N	V	A	N	A	V	R
none	27.9	24.1	06.4	08.2	10.8	16.1	04.1	06.3	05.2	01.9	02.9
correct	**37.2**	**39.1**	**39.4**	**63.3**	**56.9**	**44.1**	**64.3**	**58.3**	**57.73**	**53.4**	**44.8**
semicorrect	02.3	04.6	11.0	05.1	04.3	05.4	02.0	04.2	07.1	09.7	15.2
incorrect	25.6	18.7	24.8	13.3	15.0	22.6	22.4	17.7	19.4	13.6	17.1
> 1 selected sense	06.9	11.5	18.3	10.2	12.9	11.8	07.1	13.5	10.2	21.4	19.9
unknown	17	17	4	6	11	12	4	9	12	7	11

Table 7: Distribution (%) of the answers for the first scenario (Precision) questions for the Nouns (N), Adjectives (A), Verbs (V), Adverbs (R).

For most of the cases, all the heuristics perform well, with the users answering either correctly or semi-correctly, in particular, the *synset* heuristic that can be applied to all the POS can help the users choose between multiple senses precisely in 62.7% of the cases.

[1] when the user selected more than one sense and none were the correct one.

[2] when the user selected more than one sense and one of them was the correct one.

3.6 Discrimination Power Results

In an annotation application, it is often more important that the summary is good at discriminating between the multiple senses of the term that it applies to. Actually, the user has to choose the right concept among several choices when annotating. It is thus more important that the summary is good at discriminating between the different terms displayed than how precise it is at defining the term.

In the case of the Nouns and Verbs, we can see that the `hypernym` (*parent*) and `hyponym` (*child*) relations in WordNet provide summaries that are able to discriminate among senses quite effectively for the users (see Table 8, differences between "correct" and "incorrect" in "N" and "V" columns). However, using other terms in the same *synset* is less effective; this might be due to the fact that the other terms in the synsets are themselves quite ambiguous.

For the Adjectives and Adverbs, it is very difficult to generate a good discriminating summary and another strategy might be needed to help the user choose the right sense. For instance, it might be more informative, even if it takes more space, to show to the user an example of use of the adjective instead of a single word summary.

	Child		**Derived**	**Gloss**	**Parent**		**Similar To**	**Synset**			
Answer Type	**N**	**V**	**R**	**V**	**N**	**V**	**A**	**N**	**A**	**V**	**R**
incorrect	31.1	35.5	59.7	33.3	23.6	36.9	59.4	55.6	64.4	53.2	66.3
correct	**68.9**	**64.5**	**40.3**	**66.7**	**76.4**	**63.0**	**40.6**	**44.1**	**25.6**	**46.7**	**33.7**
unknown	10	7	29	10	11	8	4	7	11	8	14

Table 8: Distribution (%) of the Answers for the second scenario (Discriminating Power) questions for the Nouns (N), Adjectives (A), Verbs (V), Adverbs (R).

4 Related Work

While we have found a large body of research about free text summarization and paraphrasing, the problem statement is very different and the available work in that field is not relevant to our domain. Research on synonyms detection can also be found, however, here again, the approaches are far from our problem statement. To the best of our knowledge, the work on paraphrasing, while still distant, could be considered somehow similar to our approach; for instance, [6] have used WordNet for the generation of lexical paraphrases for search queries, however, in their case the paraphrases were intended for machine consumption. Similarly, [1] have used paraphrases for interactive query refinement, to assist humans during search. In the context of sense disambiguation for Semantic Web applications, [4] have used paraphrasing, but to disambiguate between different meanings of complete sentences. While, for example, Freebase or Faviki[3] use the same type of approach for helping the user choose *entities*, as far as we know, there

[3] http://www.freebase.com; http://www.faviki.com

is not yet any published work, in particular evaluations, on concept summarization to aid users during term disambiguation.

5 Conclusion

Doing fully automatic disambiguation to a sense in the underlying controlled vocabulary is difficult. Thus we need to provide users as much help as possible when they work with a controlled vocabulary. There might be many ways to ease the disambiguation task, but we believe that providing more streamlined annotation interfaces is a prerequisite. By providing a one word summary for all the senses of ambiguous words, it will be easier to display the choices of senses to the users and thus improve the manual disambiguation process.

In future work, it would be interesting to study the effect of examples of use as provided in lexical resources such as WordNet as an additional, short visual disambiguation help. Also, while we have studied the disambiguation power of the generated summaries, it would be interesting to study their direct effect on a real annotation task.

References

[1] Peter G. Anick and Suresh Tipirneni. The paraphrase search assistant: Terminological feedback for iterative information seeking. In *SIGIR*, pages 153–159, 1999.

[2] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[3] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. The MIT Press, Cambridge, MA, May 1998.

[4] Jonathan Pool and Susan M. Colowick. Syntactic disambiguation for the semantic web. In *SAAKM*, 2007.

[5] Ilya Zaihrayeu, Lei Sun, Fausto Giunchiglia, Wei Pan, Qi Ju, Mingmin Chi, and Xuanjing Huang. From web directories to ontologies: Natural language processing challenges. In *ISWC/ASWC*, pages 623–636, 2007.

[6] Ingrid Zukerman and Bhavani Raskutti. Lexical query paraphrasing for document retrieval. In *COLING*, 2002.