Multiontology Semantic Disambiguation in Unstructured Web Contexts

Jorge Gracia
IIS Department
University of Zaragoza
Zaragoza, Spain
jogracia@unizar.es

Eduardo Mena
IIS Department
University of Zaragoza
Zaragoza, Spain
emena@unizar.es

ABSTRACT

The ability of computers to automatically determine the right sense of words, according to the context where they appear, can help bridge the gap between syntax and semantics required for the full development of the Semantic Web. However, the applicability of these techniques is sometimes hampered by the unrestricted way in which humans annotate web resources, especially in folksonomies. In such cases many context words are useless (or even harmful) to determine the right meaning of another one. Furthermore, these contexts lack well-formed sentences, thus preventing syntactic analysis and other features exploited by traditional disambiguation techniques from being used.

In this paper we propose a technique for intelligent context selection, based on semantic relatedness computation, to detect the set of words that could induce an effective disambiguation. We use this technique as starting point of a disambiguation process that receives an ambiguous keyword and its context words as input, and provides a list of possible senses for the keyword, scored according to the probability of being the intended one. It combines different techniques to operate: Web-based relatedness, overlap of semantic descriptions, and frequency of use of senses. It accesses any pool of online ontologies as source of word senses, in addition to other available resources. Both our context selection technique and disambiguation method have been tested in this work, obtaining promising results.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

General Terms

Algorithms

Keywords

Semantic Web, collective knowledge, disambiguation

1. INTRODUCTION

Nowadays, there is an emergent type of web systems that constitute the so-called Social Web [20]. Such systems are centred in the participation of users, who contribute by uploading, exchanging, and tagging web content (pictures, articles, bookmarks, videos, etc.). On the other hand, the Semantic Web is described as an evolution of the current Web (consisted largely of human-readable documents) to one that includes data and information for computers to manipulate [23]. According to Gruber [11], the knowledge representation and reasoning techniques of the Semantic Web may unlock the *collective intelligence* of the Social Web, enabling a new class of applications called *collective knowledge systems*, able to generate new knowledge that are difficult to discover by other means. In order to fully accomplish the vision of collective knowledge systems, techniques to clearly define the semantics of web resources can be used to bridge the gap between the Social Web, informal and unrestricted, and the formal and structured Semantic Web. We consider Word Sense Disambiguation (WSD) as part of such techniques.

WSD techniques try to pick the most suitable sense of an ambiguous word according to the context (usually their surrounding words) in which it appears [1]. For example, the word *plant* could mean[1] "buildings for carrying on industrial labor" or "a living organism lacking the power of locomotion". It is expected that, in a text about car manufacturing, it is used in the first sense, while the second meaning may be the right one on a web page about gardening.

[1] According to WordNet 3.0 definitions (http://wordnet.princeton.edu/).

Traditional WSD techniques utilize many context features for disambiguation (part-of-speech, collocation, discourse, syntactic features, surrounding words, etc.) [1] when dealing with ambiguities in well-formed texts and sentences. Unfortunately, these features are not available in certain Web-based systems, where the context consists of unstructured bags of words, as keywords in user queries, or tags in folksonomies, for example. In fact, far from using semantic annotations, web users are becoming more and more accustomed to using tags, mainly in the context of the above mentioned Social Web. The definition of tags is a loose and implicit process where ambiguity might remain [23, 2]. Therefore, user tags provide an unstructured and highly heterogeneous context for disambiguation: usually free text, where syntactic analysis cannot be applied (as there are not well-formed sentences), and often referring to subjective impressions of users (e.g., “my favourite”, “amazing”) or technical details (e.g., “Nikon”, “photo”). Therefore, the problem of determining what are the context words that better help in the disambiguation arises, as many user tags are useless (or even harmful) for disambiguation.

In this paper, we formulate the hypothesis that the most significant words in the disambiguation context are the most highly related to the word to disambiguate. It serves as basis for the intelligent context selection technique that we propose. This context selection technique is used as the starting point in the complete disambiguation method that we also propose in this work. It enhances our previous work in the field [10], by combining 1) the use of a Web-based relatedness measure, 2) the overlap between the semantic descriptions of the word to disambiguate and the words in the context and, finally, 3) the frequency of use of senses. Our disambiguation method is not limited to the use of WordNet [17], or any other predefined ontology or lexical resource, as source of word senses. On the contrary, it dynamically exploits online ontologies, in addition to any other available resource.

In summary, our disambiguation method is intended *to provide well defined senses for ambiguous terms utilized in unstructured web contexts, expressing these senses by means of dynamically selected ontology terms.*

The rest of the paper is organized as follows. In Section 2 we study some related work. Our disambiguation method is proposed in Section 3. Section 4 contains our experimental results and, finally, conclusions and future work can be found in Section 5.

2. RELATED WORK

Techniques for explicit WSD can be classified in three groups [1], although this classification is approximate, and combined techniques are not rare (our own method falls between the first and second group):

1. *Knowledge-based* or *dictionary-based* methods, that rely on electronic dictionaries, thesauri and lexical knowledge bases, without any corpus evidence. They usually rely on semantic measures computation (e.g. [27, 22]).

2. *Unsupervised corpus-based* methods, which avoid external information and work with raw unannotated corpora. These methods usually induce word senses from training text by clustering word occurrences, classifying the new occurrences into the induced clusters/senses (e.g. [18, 13]).

3. *Supervised corpus-based* methods, that make use of annotated corpora to train from, or as seed data in bootstrapping process. They have given the best results so far, however suffering from the so-called knowledge acquisition bottleneck, a major drawback that limits their potential and scalability (e.g. [28, 16]).

The different periodic exercises carried out by Senseval initiative[2], have confirmed that the highest ranges of accuracy are only reached by *supervised* methods [25, 21]. As a matter of fact, unsupervised ones usually score below the “most frequent sense” baseline. Nevertheless, in spite of their difficulties, we advocate for unsupervised and knowledge-based methods as the most suitable ones for the Semantic Web. The knowledge acquisition bottleneck problem, which affects supervised methods, makes these methods less viable given the fast growth of semantic content currently available online and the dynamism required by emergent semantic applications.

Many unsupervised methods, however, require preprocessing tasks. For example, in [18] they construct a co-ocurrence matrix with the categories of the utilized thesaurus, populated with frequency counts from a corpus. It renders the method difficult to use with unrestricted and dynamically selected sources of senses, as we do. Also moderately supervised methods, like SenseLearner [16] result of great interest, as they increase generality with respect to purely supervised ones, while preserving a good performance. However, SenseLearner exploit certain features, such as collocation or part-of-speech, that are difficult to apply in contexts with unstructured bags of words.

A majority of traditional disambiguation methods [1, 27] rely on specific lexical resources (e.g., WordNet) or certain predefined ontologies to operate. This leads to coverage problems when dealing with words or meanings not present in WordNet. On the contrary, we search on the Semantic Web as source of word senses,

[2] Its mission is to organize and run evaluation and related activities for the semantic analysis of text. See http://www.senseval.org/

in addition to WordNet (and any other local ontology). For example, the term *developer* does not appear in WordNet 3.0 meaning "someone who develops software", but can be found in various online ontologies[3].

Regarding disambiguation methods for the Semantic Web, although their interest has been largely recognized [24], there have not been many specific efforts so far. Some remarkable exceptions are the works dedicated to entity disambiguation to enhance annotation tasks. E.g., in [12] a method is proposed to disambiguate entities in plain texts by using the background knowledge provided by populated ontologies. Other disambiguation methods have been successfully applied to ontology population, as SSI (structural semantic interconnections) [19], a general purpose disambiguator also useful for Semantic Web tasks; and SemTag [6], an application to perform automated semantic tagging of large corpora. In spite of the great interest of these works, some of them need costly preprocessing tasks (thus limiting generalization), and all of them work with a single ontology as source of senses. On the contrary, we propose in this work the use of an unrestricted pool of ontologies, to maximize the possible interpretations one can find for the words to disambiguate. Furthermore, work on disambiguation applied to the Web has not explicitly tackled, so far, the problem of selecting the more suitable words for the disambiguation context as a general and separate task, as we do.

3. DISAMBIGUATION METHOD

As mentioned above, we have found that context selection becomes a major problem when applied to certain Web-based systems where well-formed sentences are not available, as it is the case for sets of user keywords, or tags in folksonomies.

In this section we start presenting our solution for an intelligent selection of the disambiguation context when no other features than an unstructured set of words are available. This constitutes the first step of a complete disambiguation process, that determines the right sense of a given keyword, according to the context in which it appears. Our disambiguation technique reuses some well-established ideas from dictionary-based methods, however making them fully applicable to such an open context as the Web.

An earlier version of our technique is described in [10]. We have preserved here the fundamental ideas of this previous work, however adding new improvements: a preliminary context selection step, an algorithm to consider overlaps between context and semantic descriptions, and a reconsidered way of using the frequency of word senses to enhance the disambiguation result.

Our disambiguation method is not intended to substitute other well-established ones, but to be used in situations where others have difficulties to operate, for example when:

1. Dealing with unstructured contexts, as folksonomy tags, search query terms, etc. instead of well-formed texts and sentences.
2. Knowledge sources are not known in advance, and must be selected dynamically.
3. Maximizing the coverage of possible interpretations of a word (e.g., by accessing online ontologies to complement other resources such as WordNet).

Figure 1 shows the scheme of our approach (to be detailed in the rest of the section): An ambiguous keyword and its context are introduced as input. Then, a process selects the more effective context words for the disambiguation. After that, online and local resources are accessed to provide a set of candidate senses for the keyword. Finally, our disambiguation algorithm is run and the senses are weighted according to their likeliness of being the right one. For simplicity, we consider only one keyword to disambiguate, although the algorithm can be iteratively repeated when more keywords need disambiguation. The figure represents Yahoo![4] and Watson[5] as sources of web frequencies and online information respectively, but others can be used.

As motivating example, let us suppose that we want to disambiguate the word k_d =*turkey* in the following context, extracted from the set of tags that annotate a particular picture[6] in Flickr[7]):

$$C=\{\textit{roasted, perfect, poultry, son's, meat, set}\}$$

In the rest of this section, we explain how to select the most suitable words from the context to disambiguate, how to obtain candidate senses for the ambiguous word and, finally, how to deduce the most probable one.

3.1 Disambiguation Context Selection

A *keyword* k is an element of the set of strings $\mathbb{S}$ (sequences of symbols of any length over an alphabet), with a special significance. We will represent the set of all possible keywords as $\mathbb{K} \subseteq \mathbb{S}$. Let us call $C \subseteq \mathbb{K}$ the set of keywords of the disambiguation *context*, and $k_d \in \mathbb{K}$ the *target keyword* to disambiguate. In the rest of the paper, we will not further distinguish between *words* and *keywords*, that will be interchangeable.

[3] E.g., http://usefulinc.com/ns/doap#developer

[4] http://yahoo.com

[5] http://watson.kmi.open.ac.uk/WatsonWUI/

[6] http://www.flickr.com/photos/cobalt/67047133/, last accessed on 15 June 2009.

[7] A social network application to share pictures on the Web (http://www.flickr.com/).

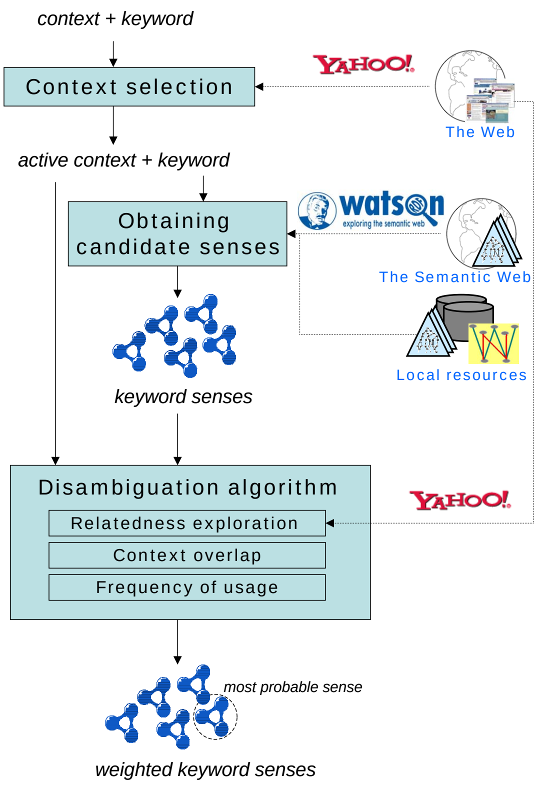

Figure 1: Scheme of the disambiguation method.

Definition 1. **Active context word.** Given a context C and a keyword to disambiguate k_d, we define active context word as any $k_i \in C$ which turns out to be significant for the disambiguation of k_d. It will be denoted as $k_i \rightarrow k_d$.

Whether a word in C is *significant* for the disambiguation of k_d or not, is something unknown in advance that should be inferred empirically after a disambiguation process, unless one applies some a priori hypothesis to decide it, as the one we will see later in this section.

Definition 2. **Active context.** Given a context C and a word to disambiguate k_d, we define active context as a $C^a \subseteq C$ such that $\forall k_i \in C^a, k_i \rightarrow k_d$.

We define *inactive context* to be the dual concept, that is, any subset of C which does not contain any active context word.

Notice that all the words in C^a are active context words, but C^a is not constrained to contain *all* the active context words in C. Therefore C can induce different active contexts, giving us the liberty of adjusting the cardinality of C^a to our necessities.

We call *semantic relatedness* the degree in which two entities (words, ontology terms, etc.) are related by any kind of semantic relationship. In order to construct the active context C^a, we rely on the following:

Hypothesis. *Given a context C and a keyword to disambiguate k_d, the likelihood of any $k_i \in C$ to be an active context word is proportional to the semantic relatedness between k_i and k_d.*

The previous hypothesis generalizes the intuition followed in [7] for the disambiguation of terms in multilingual ontology enrichment, and points out a simple mechanism for context selection, which is the relatedness computation between words. In Section 4 we confirm the validity of this supposition. This idea can be compared with the use of relatedness for disambiguation [27], but targeted to the preliminary and separate task of context selection.

Based on this hypothesis, we propose a simple mechanism to select the active context in the disambiguation of $k_d \in \mathbb{K}$ with a context $C \subseteq \mathbb{K}$: after removing repeated terms and stopterms from C, we compute a semantic relatedness $rel(k_d, k_i)$ between each context word $k_i \in C$ and the keyword to disambiguate k_d. Then we construct C^a with the context words whose relatedness score above a certain threshold. The output of this process is the active context $C^a \subseteq C$ containing the set of active context words. We limit the maximum cardinality of C^a to a certain value[8] (according to Kaplan's experiments [14], there is a number of words above which the context does not add more resolving power to the disambiguation).

This task, and some others described in the rest of this paper, require the use of a *semantic relatedness* measure. We use the relatedness measure described in [9], that computes the semantic relatedness between words, between ontology terms, or between ontology terms and words, obtaining a value between 0 and 1. It is based on elementary computations of the Cilibrasi and Vitányi's Normalized Google Distance [4], but generalized to any search engine. This measure is based on the co-occurrence of words on web pages, according to frequency counts. We use it because of its independence on the particular source of knowledge and its good performance in comparison to other traditional measures (see the evaluation we carried out in [9]). Here, and in the rest of the paper, we compute the Web-based relatedness by using Yahoo! search engine as source of web frequencies, due to its good balance between quality of results and response time [9].

[8] We use 4 in our prototype, following [14].

When computing the relatedness between *turkey* and each context word in our example, we obtain the results given in Table 1.

roasted	perfect	poultry	son	meat	set
0.30	0.21	0.29	0.19	0.29	0.19

Table 1: $rel(x, y)$ between *turkey* and the words in the context.

Therefore, the selected active context[9] is:

$$C^a = \{\textit{roasted, meat, poultry}\}$$

We clarify that our intention is to reduce the context for *disambiguation* purposes *only*. Other goals, as *data retrieval*, could still benefit of words that are irrelevant for disambiguation purposes. For example, the words "nice picture" could be relevant for a user looking for photos that others have evaluated positively. Therefore we do not discard or delete these words from the context, but only ignore them in the process of selecting the most probable sense of another word.

3.2 Obtaining Candidate Senses

In order to define the possible senses of k_d, certain semantic descriptions must be provided, which can be obtained from different sources of knowledge, such as WordNet [17], local ontologies, and pools of online ontologies (accessed by means of Watson [5] or Swoogle [8]).

The final output of this process is a set of candidate *senses*, denoted S_{k_d}, that describe the possible meanings of k_d. Each sense $s_i \in S_{k_d}$ corresponds to an ontology term (class, property or individual), or to the integration of various ontology terms of the same type, when some integration technique among terms from different ontologies is applied in order to group semantically equivalent terms (in particular, we use the technique described in [29]).

In our example, different possible senses can be obtained for *turkey*. We have accessed Watson and WordNet 2.0 to obtain candidate senses, integrating the ones that were similar enough [29]. Table 2 shows some obtained senses (they are represented, for simplicity, by their source ontologies[10] and direct parents only):

sense	source	direct hypernyms
s_1	WN + Country.daml	country, West Asia
s_2	WN + Ontosem.owl	poultry, bird
s_3	WN	unpleasant person

Table 2: First integrated senses obtained for "turkey" (summary). WN stands for WordNet.

3.3 Disambiguation Algorithm

The input to this process is the keyword k_d to disambiguate, its active context C^a, and its set of possible senses S_{k_d}. The output will be a weight for each possible sense $s_i \in S_{k_d}$, denoted $s_i|_{score}$, that represents the confidence level of being the right keyword sense according to the context.

We take into account three main contributions to compute the confidence level of each sense: 1) the web-relatedness between the sense and the words in the context, 2) the overlap between the semantic description of the sense and the words in the context, and 3) the frequency of usage of each sense.

Step 1: Web-based relatedness. First, we want to explore the semantic relatedness among the senses of k_d and the words in the context. For this task, we use the above mentioned Web-based relatedness measure between ontology terms and words [9], based on co-occurrences of terms on the Web. The idea is to establish comparisons between each sense in S_{k_d} and the words in the context, following this algorithm:

Algorithm 1 (Initial disambiguation).

```
for each sense $s_i \in S_{k_d}$ do
    for each keyword $k_j \in C^a$ do
        $r_j = rel(s_i, k_j)$
    end for
    $s_i|_{score} = \sum_j r_j / |C^a|$
end for
```

where $rel(s, k)$ measures the relatedness between sense s and keyword k.

Step 2: Context overlap. Sometimes co-occurrence of terms on the Web is not enough to conclude which sense should be activated in the disambiguation. For this reason we add to the previous computed relatedness a factor that measures the overlap between the words that appear in the context and the words that appear in the semantic definition of the sense. For this purpose we use the following algorithm:

[9] We used an empirical 0.22 as threshold in our prototype, deduced after training our system with a series of in-house experiments.

[10] The ontologies mentioned in the table can be found at http://islab.hanyang.ac.kr/damls/Country.daml and http://morpheus.cs.umbc.edu/aks1/ontosem.owl

Algorithm 2 (Adding context overlap).

```
$maxScore = max(s_1|_{score}, .., s_n|_{score})$
for each $s_i \in S_{k_d}$ do
    $signature$ = bag of words in $OC_{s_i}$
    $overlap = \frac{ComputeOverlap(signature, C^a)}{min(|signature|, |C^a|)}$
    $newScore = s_i|_{score} + (1 - maxScore) * overlap$
    $s_i|_{score} = newScore$
end for
```

where OC_s denotes the *ontological context* of s, that is, the set of ontological elements that characterize its semantics. It comprises synonyms, glosses, and labels, as well as glosses and labels of other related terms, such as (depending on the type of term) hypernyms, hyponyms, properties (also meronyms and holonyms, when using WordNet), domains, ranges, etc. In our algorithm, *signature* is an aggregation of the words that appear in the ontological context of the sense s_i. *ComputeOverlap*(A, B) is a function that returns the number of words in common, ignoring stop terms, between the bags of words A and B. The *newScore* gives 1 when the sense shows maximum relatedness as well as *total* overlap at the same time. The idea of this algorithm somehow corresponds to the Simplified Lesk algorithm [15], which studies the overlap between the gloss that describes a sense and the words in its surrounding context, and to the Banerjee and Pedersen's extended gloss overlap measure [3], We, however, exploit any information available in the ontological context.

Step 3: Frequency of usage. The *skewed frequency distribution*, or dominant use of certain senses in text collections, is a strong indication of the importance of frequency statistics for WSD [26]. It justifies considering the effect of frequency of use of senses in the disambiguation. At this point of our method, we already have a score for each sense that represents its relatedness and overlap with respect to the context words. However, sometimes these values are too close each other to clearly determine which one is the right sense. In this case neither relatedness nor overlap have been conclusive, so we add the frequency of usage of senses, if this is available in our accessed sources (e.g., WordNet).

Algorithm 3 (Adding frequency of usage).

```
$maxScore = max(s_1|_{score}, .., s_n|_{score})$
for each $s_i \in S_{k_d}$ do
    if $s_i|_{score} > proximityFactor * maxScore$ then
        $newScore = s_i|_{score} +$
                $(1 - maxScore) * normFreq(s_i)$
        $s_i|_{score} = newScore$
    end if
end for
```

where *proximityFactor* $\in [0, 1]$ indicates whether the existing scores are a clear indicator of the right sense or, on the contrary they are too close to each other and the frequency factor should be applied; and *normFreq*(s_i) is a function that retrieves a value in [0,1] proportional to the *frequency* of usage of s_i. The particular normalized frequency function we use is intended to "smooth" the possible differences of frequency counts among all senses of a word, by making them linear under a square root function:

$$normFreq(s_i) = \sqrt{a\frac{frequency_i}{\sum_j frequency_j} + b}$$

where heuristic parameters[11] a and b are constrained by $a, b \in [0, 1]$ and $a + b = 1$. The proposed formula for *newScore* in the algorithm is intended to give 1 when s_i shows maximum relatedness, maximum overlap and maximum frequency of usage at the same time.

After applying this process to the keyword senses in our example, we obtain the following weights: $s_1|_{score} = 0.23$, $s_2|_{score} = 0.93$ and $s_3|_{score} = 0.18$. Therefore, the sense that the application chooses as the most probable one for *turkey*, in the given context, is s_2, which represents its meaning as *poultry* (and corresponds with the human observation, in this case).

4. EVALUATION

We have performed an experiment conceived to validate our context selection technique, as well as our disambiguation algorithm, in a situation close to a real usage on the Web. The idea is to explore the use of ambiguous terms in web searches, particularly when looking for pictures in Flickr. As expected, many ambiguous terms can be interpreted differently depending on each obtained picture (e.g., a search of *java* in Flickr retrieves photos of "coffee beans", as well as landscapes of "the Indonesian island"). We are interested in observing the behaviour of our system when finding out the right meaning of the ambiguous search keyword, according to the context (user tags) in which it appears in each photo, in order to compare it to a human-based reference annotation.

For this purpose, we have used a corpus of 350 pictures extracted from Flickr, each one with its title and set of user tags. They were obtained from the first 25 results of different Flickr searches on the following ambiguous keywords[12] (one search per keyword): *java,*

[11] We have used $a = b = 0.5$ in our experiments (empirically inferred from previous internal tests), as well as a $proximityFactor = 0.75$.

[12] Each selected keyword has various senses in WordNet and its most frequent one was present among the retrieved set of pictures (to allow comparisons with the "most frequent

plant, star, turkey, film, arm, bush, bank, plane, clipper, nickel, mail, vessel, and *mouse.* We asked two external evaluators (both university graduates and highly skilled in English)[13] to annotate the intended meaning of the ambiguous word in each photo (exceptionally, a second possible meaning could be provided). For this purpose they considered the picture itself, its title, and the words in the set of tags. The evaluators used WordNet 2.0 as source of possible senses, but they could leave the picture untagged when they were not satisfied with any available meaning. We consider that using only WordNet as source of senses somehow limits the evaluation of our approach, which is able to operate with any ontology dynamically selected from a pool. However, we have chosen WordNet to compare with the "most frequent sense" baseline that it provides.

The inter-annotator agreement was a 79%, which is a normal value in this kind of fine grained WordNet-based experiments. Excluding the disagreements and the untagged cases, the corpus was reduced to 240 cases. The average polysemy was 4.9 senses per keyword, leading to an experimental lower bound of 20% precision (for a random disambiguation). For each test case, the input to the process was the keyword to disambiguate and, as context, the set of tags that annotate each picture. Figure 2 exemplifies one of these test cases[14], showing the results given by our system as well.

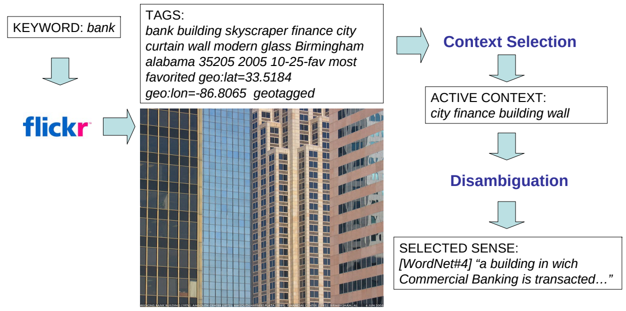
KEYWORD: *bank*
flickr
TAGS:
bank building skyscraper finance city curtain wall modern glass Birmingham alabama 35205 2005 10-25-fav most favorited geo:lat=33.5184 geo:lon=-86.8065 geotagged
Context Selection
ACTIVE CONTEXT:
city finance building wall
Disambiguation
SELECTED SENSE:
[WordNet#4] "a building in wich Commercial Banking is transacted..."

Figure 2: Example of the given results for one of the test cases.

In order to evaluate our context selection technique, we run our disambiguation algorithm with two different strategies: 1) selecting the *active context* (the most semantically related tags) as input to disambiguate, and 2) selecting as context the less semantically related tags (that we call *inactive context*). We have compared the results with respect to the human judgement and computed *precision*, which equals *recall* in this experiment since we always predict exactly one sense for each test case. Therefore, we will use the name *accuracy* to mean both precision and recall. Two baselines are provided: the use of the "most frequent sense" (MFS) given by WordNet, and a random disambiguation. The results are shown in Table 3.

	Accuracy
Active context	58±3%
Inactive context	45±3%
MFS baseline	43±3%
Random baseline	20±1%

Table 3: Averaged *accuracy* for the social tagging disambiguation experiment.

Discussion. We see that the best performance corresponds to the use of active context words in the disambiguation (58%), which outperforms an inadequate selection of context (45%). It supports our initial hypothesis, and indicates that our method for context selection behaves well. Another important conclusion is that our disambiguation algorithm, when combined with our context selection technique, beats both the random and MFS baselines in this experiment (20% and 43% accuracy respectively). This is a remarkable achievement, because the state of the art indicates that non supervised techniques rarely score above MFS baseline [25, 21, 1]. The obtained accuracy is not far from the best results obtained by supervised systems in other fine-grained experiments based on WordNet, as the English all-words track in SemEval-2007 [21] (59%, with 51% MFS baseline) and Senseval-3 [25] (65%, with 61% MFS baseline). This comparison is, of course, merely indicative (due to the different nature of the experiments), but confirms that we advance in the right direction.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have explored how to tackle disambiguation in Web-based applications, solving the problem of context selection when dealing with unstructured and heterogeneous contextual information. Our main contribution is twofold:

1. A method for intelligent selection of disambiguation context, based on the hypothesis that the relevance of a context word for disambiguation is proportional to its semantic relatedness with the word to disambiguate.

2. A disambiguation method based on: computing Web-based relatedness measures, overlap between context words and semantic descriptions, and word sense frequencies. It does not rely on any particular ontology or lexical resource to operate, be-

sense baseline").

[13] Here we follow the guidelines of disambiguation benchmarks [25, 21] that commonly use two human experts to create reference semantic annotations.

[14] http://www.flickr.com/photos/dystopos/18396625/, last accessed on 14th April 2009.

ing appropriate when knowledge sources are not known in advance.

Our evaluation, in the rather difficult context of social tagging, has shown the validity of our hypothesis for active context selection. In fact, we observed a significant improvement of the disambiguation performance when this context selection technique is applied. Furthermore, we have found that, in our experiment, our disambiguation algorithm beats the "most frequent sense" baseline, which is a major achievement for a non supervised method.

As future work, we will study particular applications of our techniques to enhance other semantic web systems, such as semantic enrichment of folksonomies, multilingual ontology enrichment, or semantic annotation of web content. Also a more clever way to auto-adjust the heuristic parameters we use in our methods will be explored.

Acknowledgments. This work is supported by the CICYT project TIN2007-68091-C02-02.

6. REFERENCES

[1] E. Agirre and P. Edmonds. *Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] S. Angeletou, M. Sabou, and E. Motta. Semantically enriching folksonomies with FLOR. In *Proc. of 1st International Workshop on Collective Semantics: Collective Intelligence and the Semantic Web (CISWeb 2008) at ESWC'08*, June 2008.

[3] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, August 2003.

[4] R. L. Cilibrasi and P. M. Vitányi. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, March 2007.

[5] M. d'Aquin, C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, and E. Motta. Characterizing knowledge on the semantic web with Watson. In *5th International EON Workshop, at ISWC'07, Busan, Korea*, 2007.

[6] S. Dill, N. Eiron, D. Gibson, D. Gruhl, and R. Guha. Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *WWW'03*. ACM, 2003.

[7] M. Espinoza, A. Gómez-Pérez, and E. Mena. Enriching an ontology with multilingual information. In *Proc. of 5th European Semantic Web Conference (ESWC'08), Tenerife (Spain)*, pages 333–347. Springer Verlag LNCS, June 2008.

[8] T. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng. Swoogle: Searching for knowledge on the Semantic Web. In *AAAI 05 (intelligent systems demo)*, July 2005.

[9] J. Gracia and E. Mena. Web-based measure of semantic relatedness. In *Proc. of 9th International Conference on Web Information Systems Engineering (WISE'08), Auckland, New Zealand*. Springer Verlag LNCS, September 2008.

[10] J. Gracia, R. Trillo, M. Espinoza, and E. Mena. Querying the web: A multiontology disambiguation method. In *Sixth International Conference on Web Engineering (ICWE'06), Palo Alto (California, USA)*. ACM, July 2006.

[11] T. Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):4–13, july 2007.

[12] J. Hassell, B. Aleman-Meza, and I. B. Arpinar. Ontology-driven automatic entity disambiguation in unstructured text. In *International Semantic Web Conference*, 2006.

[13] R. Ion and D. Tufiş. RACAI: Meaning affinity models. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[14] A. Kaplan. An experimental study of ambiguity and context. *Mechanical Translation*, 2(2):39–46, 1955.

[15] A. Kilgarriff and J. Rosenzweig. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1-2), 2000.

[16] R. Mihalcea and A. Csomai. SenseLearner: word sense disambiguation for all words in unrestricted text. In *Proc. of 43nd Annual Meeting of the ACL (ACL'05)*, Morristown, NJ, USA, June 2005. Association for Computational Linguistics.

[17] G. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), November 1995.

[18] S. Mohammad, G. Hirst, and P. Resnik. Tor, TorMd: Distributional profiles of concepts for unsupervised word sense disambiguation. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[19] R. Navigli and P. Velardi. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1075–1086, 2005.

[20] T. O'Reilly. What is web 2.0. design patterns and business models for the next generation of software, September 2005. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

[21] S. Pradhan, E. Loper, D. Dligach, and M. Palmer. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[22] P. Resnik. Disambiguating noun groupings with respect to Wordnet senses. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 54–68. Association for Computational Linguistics, 1995.

[23] N. Shadbolt, T. Berners-Lee, and W. Hall. The semantic web revisited. *Intelligent Systems, IEEE*, 21(3):96–101, 2006.

[24] A. P. Sheth, C. Ramakrishnan, and C. Thomas. Semantics for the semantic web: The implicit, the formal and the powerful. *Int. J. Semantic Web Inf. Syst.*, 2005.

[25] B. Snyder and M. Palmer. The English all-words task. In *ACL 2004 Senseval-3 Workshop*, Barcelona, Spain, July 2004.

[26] C. Stokoe, M. P. Oakes, and J. Tait. Word sense disambiguation in information retrieval revisited. In *Proc. of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'03)*, 2003.

[27] S. P. Ted Pedersen, Satanjeev Banerjee. Maximizing semantic relatedness to perform word sense disambiguation, 2005. University of Minnesota Supercomputing Institute Research Report UMSI 2005/25.

[28] S. Tratz, A. Sanfilippo, M. Gregory, A. Chappell, C. Posse, and P. Whitney. PNNL: A supervised maximum entropy approach to word sense disambiguation. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[29] R. Trillo, J. Gracia, M. Espinoza, and E. Mena. Discovering the semantics of user keywords. *Journal on Universal Computer Science (JUCS). Special Issue: Ontologies and their Applications*, 13(12):1908–1935, December 2007.