Journal of Engg. Research **Vol. 5 No. (2) June 2017 pp.** *1-18*

إنتاج الروابط المفقودة للعلاقات الدلالية في ويكشنـري

عبدالله باواكد
جامعة جدة ، جدة ، المملكة العربية السعودية

الخـلاصـة

في كثير من الحالات، تحمل الكلمة الواحدة العديد من المعاني. تقدم ويكشنري وسيلة لعرض معاني الكثير من الكلمات المختلفة بالإضافة للعلاقات الدلالية الخاصة بها. ومع ذلك، فإن ويكشنري في شكله الحالي، يحتوي على علاقات الدلالية تربط الكلمات ربطاً سطحياً لا يشمل المعاني الخاصة بهذه الكلمات. في هذا البحث، نقترح طريقة جديدة لتوليد نوع جديد من الروابط للعلاقات الدلالية داخل ويكشنري. الطريقة تعتمد بشكل أساسي على ميزات سطحية تعتمد فقط على هيكل ومحتوى ويكشنري لربط معاني الكلمات بالعلاقات الدلالية بدون مساعدة من أي قواعد معجمية أو قواعد معرفة الخارجية. نقدم تفاصيل الطريقة التي استحدثناها وكيفية التنفيذ. بالإضافة إلى ذلك، قمنا بعمل تقييمات للنظام المستحدث ونعرض في هذا البحث النتائج المشجعة التي تحصلنا عليها للنظام، خاصة عند مقارنته بأنظمة أخرى حديثة من دون الحاجة إلى الوصول إلى موارد خارجية أو بيانات للتدريب إضافية.

Generating the missing links for semantic relations within Wiktionary

Abdullah Bawakid

University of Jeddah, Jeddah, Saudi Arabia

abawakid@uj.edu.sa

ABSTRACT

In many cases, a single presentation of a term may carry multiple meanings. Wiktionary provides a way for viewing the meanings of the different terms it stores in the form of senses. It also provides semantic relations. However, Wiktionary, in its current form, contains semantic relations linking Wiktionary entries at the term level. Links for semantic relations connecting entries at the word sense level do not currently exist in Wiktionary. In this paper, we propose a novel method for generating a new type of links for semantic relations within Wiktionary. This is effectively applied to aligning the source words senses for semantic relations in Wiktionary with their corresponding target word senses. We use surface-level features that rely only on the structure and content of Wiktionary for completing this task without the aid of any external lexical or knowledge bases. We present the details of the method and how it was implemented. Additionally, we describe the evaluations that we performed and illustrate the competitive results we obtained, especially when compared to other systems. Our findings indicate that our system outperforms the baselines and performs similar to state-of-art systems without requiring access to external online resources or training data to run.

Keywords: Semantic Relations, Aligning Words Senses, WSD, Wiktionary

INTRODUCTION

With the continuous expansion of the web and the data it contains, the need for advanced measures to handle and process this data has been intensifying. Work in the area of Natural Language Processing (NLP) has attempted to address this challenge from different perspectives based on the task and application at hand. For instance, research work on multiple tasks in NLP such as automatic questions answering, text summarization, documents classifications and words sense disambiguation has often involved the usage of external aiding resources in addition to statistical and training-based methods.

The external aiding resources that were used in the literature have been of different types and sizes. They have ranged from electronic dictionaries to lexical and semantic knowledge bases. Traditional systems have relied on electronic dictionaries which are simply machine-readable dictionaries (Krizhanovsky and Smirnov, 2013). These dictionaries generally provide definitions for terms in one or more languages but do not contain direct connections between the language terms. Lexical and semantic ontologies on the other hand provide explicitly mentioned relations and links between the different terms contained within the ontology. These links can be very useful for the logical inference of knowledge that may not be explicitly mentioned in the text being processed (Garrod and Terras, 2000).

Construction of lexical and semantic knowledge bases have been attempted several times in the literature. Some knowledge bases were constructed manually by language experts in specific languages such as WordNet (Miller, 1995) and GermaNet (Kunze and Lemnitzer, 2002). Others were automatically compiled from knowledge base resources that were collaboratively built by online users but were not meant to be machine readable from the start such as Wikipedia (Hoffart et al., 2013). While the quality of manually built resources usually is superior, their size is considered limited especially when compared to the automatically constructed resources.

In this paper, we examine Wiktionary and how it can be better prepared for usage in the domain of NLP. Wiktionary is an online lexicon resource that is constructed and managed by a large number of web users. Because of its active and large community, its content is usually fast to update in case errors are detected or maintenance is required. Hence, it has a considerably good quality and coverage in addition to being large in size. It also has the advantage of not focusing only on concepts as is the case with Wikipedia, but also on terms of different parts of speech including verbs, adverbs and adjectives. This makes it particularly useful for specific NLP applications that require processing not only concepts but also verbs and the relationships they convey.

Wiktionary, in its current form, lists semantic relations with generic links connecting entries at the term level. Links for semantic relations connecting entries at the word sense level do not currently exist in Wiktionary. For example, consider the Wiktionary page for the term accept which is shown in Fig. 1. One can note the antonym semantic relation linking the term accept to another term, namely reject. We can also see a snippet from the reject Wiktionary page in the same figure showing two possible meanings (or senses) for the term reject: the first being "to refuse to accept" while the second is "to block a shot". A human, using his English language perception and background knowledge, would be able to tell that the first sense of reject is the one referred to for the antonym semantic relation (accept, reject). However, machines would require a method for coming up with a similar conclusion. In this paper, we attempt to solve this problem by providing a novel algorithm that aims to align the source word senses for the semantic relations within Wiktionary to their corresponding target word senses.

The main contribution of this paper is as follow: (i) We present a novel method for generating a new type of links for semantic relations within Wiktionary that connect words pairs at the sense level; (ii) We illustrate the effectiveness of our method through the evaluation that was performed along with the obtained results. When compared to other systems that were implemented in the literature, the results we obtained were competitive especially when considering that our method relies mainly on surface-level features from Wiktionary and requires no training or access to external online resources. It is therefore applicable to any of the 172 languages contained within Wiktionary, even those which are under-resourced.

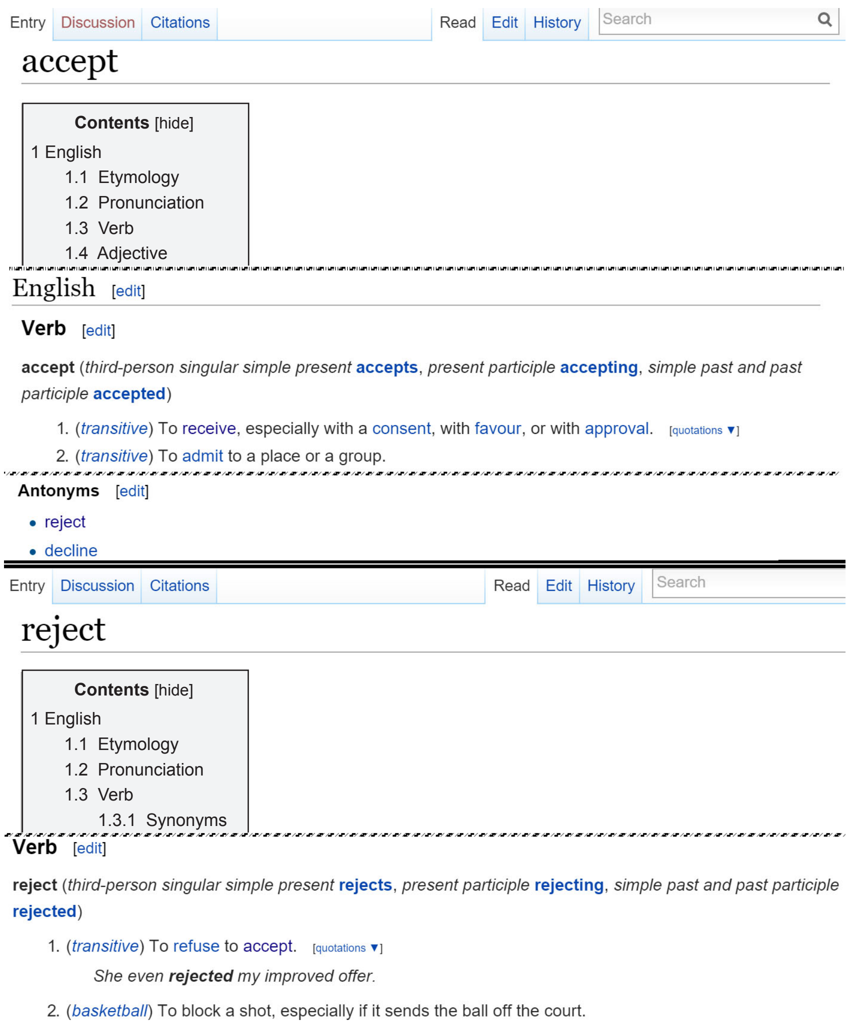

Fig. 1: Antonym semantic relation for the words pair (accept, reject) and the two possible senses for reject

In the next section we give a background on the most common lexical and semantic knowledge bases, and related work. In the following section we describe how our algorithm was implemented and evaluated. We also report the evaluation results we obtained and how the system compares with others. Finally, in the last section we give our conclusion and explore potential future work.

BACKGROUND AND RELATED WORK

Over the past two decades, one can note an increasing transition from the usage of electronic dictionaries to the usage of lexical and semantic knowledge resources in the domain of NLP. An electronic dictionary provides similar content to that of a paper dictionary but in a machine-readable form. Among the earliest work that utilized e-dictionaries is that of (Leffa, 1992) in which several studies were conducted to illustrate their importance for making a second language easier to comprehend especially for foreign students. In (Pirkola et al., 2001), e-dictionaries were successfully used to construct a query-based model on four language pairs for information retrieval. Even though e-dictionaries were found to be useful in the past, they were also lacking especially when compared with lexical and semantic resources. In addition to giving the meanings of different terms, the lexical and semantic resources may provide links between the meanings of the different terms. This can be useful for deducing logical connections between different types of text fragments (Pustejovsky, 2012) as well as the logical inference of knowledge that is not necessarily present in the processed source documents (Garrod and Terras, 2000).

One of the most common lexical and semantic resources in the domain of NLP is WordNet. It was constructed by expert linguists in a controlled manner to ensure its accuracy and quality (Miller, 1995). It is generic in the sense that it delivers the meanings of a large number of words in the English language from different domains. In addition to listing the words meanings, it provides connections between the meanings of terms themselves (Peters and Peters, 2000). WordNet was used in the domain of NLP for many applications such as text categorization (Li et al., 2012), text clustering (Bouras and Tsogkas, 2012), sentiment classification (Balamurali et al., 2011) and question answering (Clark et al., 2008). Among the main shortcomings of expert-made resources such as WordNet is that it is not always up to date as making continuous updates by a few expert linguists is both a slow and costly process. This also makes their coverage limited, especially when compared to the other alternatives.

An alternative to using expert-made ontologies has come about with the emergence of the internet encyclopedia, Wikipedia. Web volunteers constructed Wikipedia in a collaborative manner rendering it the largest encyclopedia to date. Because it has an active community, it is continuously updated and its coverage is also considerably larger than expert-made lexical and semantic resources. Each article in Wikipedia provides information and details about a specific concept. The research community used it mainly for detecting concepts within text and deducing relationships between the detected concepts. For instance, in (Witten and Milne, 2008) the authors suggested a method for computing the semantic relatedness between Wikipedia concepts using the inner and inter links of Wikipedia. In (Gabrilovich and Markovitch, 2007), the semantic relatedness was computed using a method named Explicit Sematic Analysis which relies on the textual content of each article. After creating a method for computing the semantic relatedness between concepts, a semantic knowledge resource is formed which can be utilized in several NLP tasks such as sentiment analysis (Mukherjee and Bhattacharyya, 2012), text documents summarization (Nastase, 2008) and text simplification (De Belder and Moens, 2010).

Similar to Wikipedia, Wiktionary was also constructed by web volunteers, but is tailored to focus more on the lexical terms of each language. In a way, it is similar to a dictionary since it attempts to include a separate article for each individual term in each language. However, it also includes other semantic and lexical information including the meaning of each term, its parts of speech, pronunciation, semantic relations, and translations. In contrast to Wikipedia, it does not provide details about concepts, but rather on the different terms in each language along with their meanings. In addition to nouns, it provides details about verbs, adjectives, adverbs and the semantic relationship between them, too.

Wiktionary cannot be used as is in NLP-based application. Just like Wikipedia, it needs to be pre-processed and prepared before it can be utilized in different NLP-related applications. Among the most common research efforts in the domain of NLP that utilizes Wiktionary and Wikipedia is the project of BabelNet (Navigli and Ponzetto, 2012). BabelNet is a multilingual semantic network that attempts to combine WordNet with other resources such as Wikipedia and recently Wiktionary. It attempts to identify the named entities and concepts from different resources and matches them with their suitable counterparts in WordNet. It also attempts to disambiguate concepts while linking entities from the different resources with the assumption that a context for the ambiguous sense is always present. Another prominent project in the field is DBnary (Sérasset, 2012; Sérasset and Tchechmedjiev, 2014) which is a multilingual linked data store constructed from Wiktionary. It provides lexical chains connecting word entries from different languages to each other. We focus in this paper on sense-level connections, specifically between the semantic relationships provided in Wiktionary.

Among the challenges that need to be handled in Wiktionary is the semantic relations links that connect entries at the term-level. Connections between the senses of the terms for semantic relations do not exist in Wiktionary. This problem is the focus of this paper and it bears similarity with the task of automatic Word Sense Disambiguation. With Word Sense Disambiguation, an attempt is made to disambiguate a word given the context it appears in while with Wiktionary we aim to align source words senses for semantic relations with their suitable target word senses.

Several methods were proposed in the literature that aim at disambiguating words senses. Some approaches used machine learning for this task. In general, they work by constructing a classifier with co-occurrence resolution features for assigning senses to their matches. They usually require annotated training data and may employ various statistical and probabilistic models such as Neural Networks and Decision Lists and Trees. Examples for algorithms employing such approaches are described in details in (Pedersen, 2000), (Komiya and Okumura, 2012) and (Wang and Hirst, 2014). Other systems employ methods that rely on lexical and semantic knowledge resources. They attempt to utilize the resources features which includes its structure, textual content and links for selecting the most suitable sense for an ambiguous word. The system described in (Banerjee and Pedersen, 2002) employed WordNet for this task. Wikipedia was also employed in many other systems as in (Dandala et al., 2013) and (Li et al., 2013).

The method we propose in this paper relies on Wiktionary. It differs from the above mentioned methods in that it relies on a limited number of surface-level features from Wiktionary with no training that is required. Some of the features we employ such as the definition similarity feature are similar to those used in the system of (Dandala et al., 2013) with Wikipedia, but we adapted them to be used with Wiktionary in a different task than theirs. In addition, the collective set of features we employ is unique and collectively it differs from all those that were explored in the literature.

It should be noted that the definition similarity feature we employ was also used in the literature for different and related tasks. For instance, the work of (Niemann and Gurevych, 2011) focused on the alignment of WordNet synsets to their corresponding Wikipedia Concepts. Their method primarily used the Personalized PageRank (PPR) (Haveliwala, 2002) algorithm for computing the definition similarity between WordNet Synsets definitions and the inner content of Wikipedia articles. The work of (Pilehvar and Navigli, 2014) also targeted a similar but a more generic task, namely the automatic alignment of heterogeneous lexical resources including Wikipedia, WordNet and Wiktionary. Their definition similarity computation also involved the usage of PPR. In our work, we use the definition similarity for linking Wiktionary senses with other Wiktionary senses as per the provided semantic relations within Wiktionary. We generate the definition similarity by computing the normalized dot product for the representative vectors for each compared sense as detailed in the methodology section below.

The work proposed in (Meyer and Gurevych, 2010) addresses the same problem we are addressing in this paper but it uses mainly Explicit Semantic Analysis. In the evaluation section we compare our method with that of ESA's and provide the performance results of each. More recently, an effort was made in (McCrae et al., 2012) to also tackle the same problem while constructing a resource that combines both WordNet and Wiktionary. Their method relied mainly on computing the Edit Distance (Navarro, 2001) between the definitions of the compared words for each semantic relation. In the implementation of their algorithm, they assumed based on their observation, that similar order of the definitions between different sources is often the case and that glosses of definitions from the source contains (or is contained) the gloss of the target sense. We also compare the performance of our method to theirs in the evaluation section.

METHODOLOGY

We describe in this section the methodology we employed for generating sense-level links for semantic relations within Wiktionary. When implementing our method, we downloaded the latest Wiktionary dump from the main Wikimedia dump website1. We used the open-source Java Wiktionary library provided by the authors in (Meyer, 2013) as a basis for parsing Wiktionary and extracting the required features. Our main focus was directed towards the English language of Wiktionary. However, the methodology we describe can be adapted and applied to any other language that exist in Wiktionary. We start by going through the features we rely on in our system. Then, we explain how the features are utilized to match words senses in Wiktionary.

Features

For each word w in Wiktionary, there exists a number of senses that define the meaning of that word based on its part of speech. For instance, it can be noted from Fig. 1 that there exists two parts of speech for the term accept, namely Verb and Adjective. The verb version has nine possible senses while the adjective has a single sense. In Wiktionary, a word may also have one or more semantic relations such as antonyms, synonyms, hypernyms and hyponyms. Each semantic relation links a particular sense of w to a target word t without specifying the sense of t. An example for this was also presented in Fig. 1 for the antonym semantic relation linking the source term accept with the target term reject. The problem we attempt to solve is to identify for a given semantic relation r which of the senses of the target term t is the best representative sense for the given sense of the source term s. Based on our study to different Wiktionary entries containing different types of relations and the analysis we performed, we selected the following features which are integrated in our algorithm to disambiguate the senses being linked to by the semantic relations in Wiktionary.

Semantic relations

Wiktionary has different types of semantic relations defined and used to link many of its terms. They are generally used to indicate a level of semantic relatedness between a term and another. Each relation consists of a pair of words typically referred to as source term s and target term t, and is written in the form r(source, target) to indicate the relation direction (Jurafsky and Martin, 2000). A source in Wiktionary is usually associated with a particular sense of the term while the target is not. An example for a relation is the synonym relation (drive, make) represented in Wiktionary in the form (cause to become): make under the details page for the word drive. It is possible to distinguish the drive sense referred to by the above relation through examining the encoded phrase within the parenthesis (cause to become) and matching it with the corresponding sense as illustrated in Fig. 2.

Some of the semantic relations defined within Wiktionary are synonymy and antonymy. For these two relations, it can be inferred that a relationship between any words pairs is mutual. For instance, the synonym relation (beautiful, attractive) can also derive another synonym relation (attractive, beautiful). This derived relation can be useful in identifying which sense of the word beautiful is a synonym to what sense of attractive. Other types of semantic relations that exist within Wiktionary include hyperymy, hyponymy, meronymy and holonymy. For this set of relations an inverse relation can be derived in the form of meronymy against holonymy and hyponymy against hypernym. For instance, the hyponym relation (animal, mammal) derives the hypernym relation (mammal, animal).

Fig. 2: Synonym relation for the words pair (drive, make)

Definitions similarity

For this feature, we examine the definition of the source word sense in addition to the definitions of each of the possible target word senses tj for the relation at hand. We noted that in many cases having a number of shared terms between the definition of a term and its corresponding relation sense may indicate that this sense is the right one for this relation. This is encoded in our work by forming representative terms vectors, after applying stemming and removing stop words, for s and the senses tj and then computing the normalized dot product to determine their similarity. We also take into account the existence of the source word into the definitions of the senses tj and give it double the weight to highlight its significance. An example for this is presented for the word ‘page’ which has ‘side’ as one of its synonyms as illustrated in Fig. 3. It can be noted that the fifth sense of ‘side’ is the most suitable sense for the given relation.

Fig. 3: Synonym relation for the words pair (page,side) illustrating the Definition Similarity feature

The implementation for computing the definition similarity between two senses in our system involved forming a vector of words for each sense a and b. The words are simply taken from the definition provided for each sense in Wiktionary. After removing stop words and stemming all the terms in the sense definition, each vector would contain the frequency of the terms in the sense definition. After forming the vectors, we generate a similarity score by computing the normalized dot product as follow:

$$TxtSim(a,b) = \frac{a \cdot b}{||a|| \times ||b||} \qquad \text{where} \quad ||a|| = \sqrt{\sum_{i=1}^{N} a_i^2}$$

Hence, the greater the obtained score, the more similar the two compared senses are. We use this feature whenever applicable (if the definition is available for the two senses compared) as outlined in the methodology section below.

Shared labels

The definition of many senses in Wiktionary is preceded by context labels. These labels often provide links to categories that are directly related to the word sense. They aid in distinguishing a sense from the rest of the word senses. We noted that these labels can also help in linking senses between different terms in Wiktionary. Additionally, there is a hierarchical categories structure within Wiktionary. Nodes in this structure include both domain and lexical categories names with the edges being links among them. For instance, the category en:Socialism is the parent for the categories: en:Communism and en:Marxism. It is also a child of the en:Ideologies category. An English term that belongs to a specific category would have its category name at the bottom of the Wiktionary page in the format en:categoryName. Take the words 'Air' and 'Aria' as an example. The former has a sense defined as "(music) A song, especially a solo; an aria" which belongs to the Music category. The term 'Aria' also has one of its senses defined as "(music) aria, song". Since the two mentioned senses share the parent category Music, they have a higher probability of being related than the rest of their senses.

In addition to considering terms pairs that share the same category, we also consider similarities in the parent categories for one or both terms. This can especially be useful for categories which are very specific and contain only a handful number of terms. Take the terms 'water' and 'liquid' as an example where a sense for the first belongs to the Water category while the second belongs to the Liquid category which has the Water category as one of its child categories.

Most common sense

It is often the case that senses are arranged in Wiktionary based on how common their usage are for each term. Hence, one of the features we consider in our method is selecting the first ordered sense as a top candidate for the word at hand. In addition, there are many words in Wiktionary which contain only a single sense. In this case, we assume that the only available sense is the correct one especially when dealing with relations. For instance, the word animal has a hyponym relation with the word frugivore which has a single sense.

Aligning words senses

We begin by forming a set of possible senses for the target term t. If the term t has only one sense defined in Wiktionary, then that sense would be its representative candidate. If, however, more than one sense exists, we compute a score for each sense of t. The higher the score, the more related the sense is to s. The sense with the highest score would then be chosen as the best representative sense for s given the relation r.

Computing a score for each sense involves the usage of the previously defined features. A pseudocode for the algorithm we implemented is illustrated in Fig. 4. We apply the Definitions Similarity features with the normalized dot product of the terms vectors for each possible sense e and the source s. The shared labels are considered with the aid of the function Lbls which also considers the category hierarchy of the labels within Wiktionary to include matches of parent categories with categories as explained previously. The semantic relations within Wiktionary for both the target and the source terms are considered with the function SRelation which checks for symmetric or inverse relations between the candidate senses and the source term sense s. The score for each sense is then computed and stored in CandidatesScores.

The weights that were assigned to the three main features TxtSim, ShrdLbls, and SemRel are 10, 100, 1000. These weights were chosen to reflect the importance, coverage and precedence of the mentioned features. The feature that has the least coverage within Wiktionary is that of ShrdLbls. The difference in the precision of TxtSim and SemRel is very small but TxtSim has more than 52% coverage than SemRel. We provide more details about the coverage and precision for each of these features later in the evaluation section. However, it should be emphasized that the above weight numbers were selected to reflect which features overtake the other features if they all exist in the compared senses pair (which is not always the case by the way). For instance, if a pair of senses contains a definition for the target sense as well as the source sense and there are shared labels between the two sense, the feature that would have an actual effect on the total score is ShrdLbls. If the two compared senses do not have any shared labels but there exists a definition for both the target and the source sense, the computed score would be affected by the TxtSim feature the most. In a similar way, the feature of MCS would only have an effect on the total score if the two compared senses lack the other three features in Wiktionary.

Note that the features we employ may not all be applicable to some words in Wiktionary. Some words for example may not have any semantic relations. In that case, the score for SemRel would then be zero which effectively leads to ignoring this feature and considering the rest of the applicable features. It should also be noted that we give priority to some features over the rest. For instance, SemRel would have the most effect on a sense score if it is available, followed by ShrdLbls and then TxtSim. The feature computed in MCS would only have a real effect on promoting a sense in the candidate list if the rest of the features were not possible to compute due to their unavailability.

EVALUATION

To evaluate our system, we used the dataset that was manually constructed with the aid of two human annotators in (Meyer and Gurevych, 2010) by collecting semantic relations randomly from the German version of Wiktionary. Each human annotator would then select the most suitable sense for the target term of each semantic relation. The dataset provides the selection of each human annotator along with their selection agreements and disagreements. Each of the possible target senses was tagged by the human annotator with a plus (+) if the semantic relation for the senses pair holds, or a negative (-) if it does not. The human annotators were allowed to obtain consultation from external resources such as online encyclopedia, dictionaries, etc. However, they were not allowed to consult each other. The constructed dataset consists of 920 possible target senses for 250 semantic relations. The usage of this public dataset, which was entirely prepared and built by the authors of (Meyer and Gurevych, 2010), in our evaluation allowed us to compare the results of our system against other state-of-the-art systems that also utilized the same dataset for similar problems.

```
function GetMSS(r,s,t)
 Candidates ← AllSenses(t)
 CandidatesScores ← {0}
 if( sizeOf(Candidates) = 1 )
  ChosenSense ← Candidates[0]
 else
   for each sense e in Candidates do
    TxtSim ← ( defV(e) • defV(s) )
    ShrdLbls ← #( Lbls(e) ∩ Lbls(s) )
    SemRel ← ( e ∈ SRelation(s) ) ? 1 : 0
    MCS ← isFirstSense(e) ? 1 : 0
    CandidatesScores(e) ← 10 × TxtSim +  100 × ShrdLbls +
                            1000 × SemRel + MCS
   end
  ChosenSense ← argMax c ∈ Candidates CandidatesScores(c)
 end
 return ChosenSense
end
```

We summarize the results we obtained from the performed evaluation in Table 1. The original evaluation in (Meyer and Gurevych, 2010) involved comparing the output of their system with the selections made by each of the two human raters. We do the same by treating the dataset as 920 binary annotations that need to be evaluated against human annotator 1 and human annotator 2. We report the results with the observed agreement Ao along with the kappa statistics κ (Artstein and Poesio, 2008) for the chance-correlated agreement. The results obtained with our method is labelled with WikMSS. We also present the results obtained by ESA which refers to the method used by the authors who prepared the dataset (Meyer and Gurevych, 2010). FirstSense refers to a baseline we used which marks the first sense in any list of candidate senses for a target term as the chosen sense. HumanRater serves as an upper bound and refers to inter-rater agreement on the chosen senses. Levenshtein refers to our implementation of the edit distance method applied by the authors in (McCrae et al., 2012) to a similar problem to the one we are tackling.

Table 1: Evaluation results using dataset1

	A_o-1	A_o-2	*κ-1*	*κ-2*
WikMSS	**0.82**	**0.84**	**0.58**	**0.64**
ESA	0.79	0.82	0.48	0.57
Levenshtein	0.70	0.72	0.39	0.48
FirstSense	0.78	0.79	0.45	0.50
HumanRater	0.89	0.89	0.73	0.73

It can be noted from the obtained results that our methodology gave a competitive performance when compared with others. Our system agreed with the first human rater on 0.82 of the cases, and with the second human rater on 0.84 of the cases. The obtained κ statistics for our system also suggest good reliability especially when compared with the baseline and the rest of the systems. The method of ESA relies exclusively on computing the semantic relatedness between senses definitions for the disambiguation tasks. Some target words senses may contain short or no definition text, rendering the disambiguation task with ESA alone difficult to achieve. With our method, we employ different sets of features that not only rely on the senses definitions, but also on their labels and categories. This allowed our system to perform the disambiguation using different sets of aspects within Wiktionary that may not all be available for all words senses all the time. Another major difference is that in contrast to ESA, our method does not require the usage of external knowledge repositories such as Wikipeida. It employs only surface-level features that exist in Wiktionary in addition to its structure, making it applicable to many different languages regardless of how much background resources exist for that language.

The Levenshtein method was implemented by computing the edit distance between the glosses of the target term senses against the gloss of the source term. If the gloss for one of the target senses is a substring of (or contains) the gloss of the target term sense, then this target sense would be a match. Otherwise, the senses pair with minimum gloss edit distance would be selected. We notice that Levenshtein obtains the lowest scores for all metrics as it relies exclusively on the existence of glosses. ESA also requires the presence of glosses but it is enhanced by background knowledge from Wikipedia to determine the relatedness between terms in different glosses.

In order to obtain an even better view of the performance of our system on the English Wiktionary, we used another larger dataset that was also manually built by human annotators in (Meyer and Gurevych, 2012). The dataset was constructed in a similar manner to that of the previous dataset as semantic relations were randomly selected from Wiktionary and human annotators were asked to decide whether agreement between a source sense and target senses for a semantic relation holds or not. The dataset consists of 1117 possible target senses for 394 English semantic relations. In addition to being larger than the previously used dataset, it differs from it in that the sampling of semantic relations from Wiktionary is more balanced. Semantic relations were selected according to their part-of-speech, type and number of target senses. This helps in avoiding strong bias against terms of particular part of speech such as the synonyms of nouns.

The performance of our system with the second dataset is displayed in Table 2. The reported performance uses the accuracy (Acc), precision (Prec), recall (Rec) and F1 (Manning and Schütze, 1999) measures. They are the same measures that were utilized for evaluating the WKTWSD system described in details in (Meyer and Gurevych, 2012) which was implemented by the second dataset creators. With WKTWSD, the authors utilized a larger set of features extracted from Wiktionary for the same problem we are tackling. However, the exact set of features they employed as well as the algorithm implemented differ from ours. For instance, they employed an external resource, namely the Bing Translator made by Microsoft, to compare the gloss of senses for terms from different languages. The translation serves as a bridge to extend those senses which lack a definition within Wiktionary. Another difference is that our algorithm utilizes the categories structure within Wiktionary while their system does not. We also report the result of running a training-based method, namely the J48 decision tree, on the same dataset. J48 is provided as part of the Weka data mining toolkit (Hall et al., 2009) and is an implementation of the C4.5 algorithm (Quinlan, 1993, p. 5). Training and evaluation for J48 was conducted by the authors of dataset2. We report their findings in Table 2 for the sake of comparison. The Weka toolkit was used for the training of J48 with a five-fold cross validation. No attempt was made to fine tune the training parameters in order to avoid overfitting the setup with the given dataset. As a lower baseline, we show the result of running a basic system, FirstSense, that simply selects the first target sense as the most suitable.

Table 2: Evaluation results using dataset2

	Acc	Prec	Rec	F_1
FirstSense	0.81	0.75	0.74	0.74
WikMSS	0.83	0.79	**0.81**	**0.79**
WKTWSD	**0.84**	0.78	0.80	**0.79**
J48	0.83	**0.81**	0.71	0.76
Human	0.91			0.89

The reported results with the second evaluation illustrates that our system gives superior results against the baseline. It also performs competitively against WKTWSD. While the F1 score is the same for both, precision and recall for our system is better than theirs. However, the accuracy of their system exceeds ours. While the differences in all reported scores between the two systems are marginal, it should be emphasized that our method is lighter in the sense that it depends on a smaller number of features than theirs. In addition, the algorithm we employ does not require the usage of any external online resources such as the Bing Translator. This makes it applicable to a broader set of language which lack extensive online resources. The reported results for J48 is also comparable with ours. This indicate that a training-based system does not necessarily give superior performance by a higher margin than other systems that require no training for the problem we have at hand. They usually give higher precision but at the cost of lower recall. This is mainly caused by the varying number of possible target senses for the different semantic relations within Wiktionary. While some relations may have a single target sense, others may have ten or more target senses which leads the classifier to identify a larger number of true negatives against true positives. This is a similar finding to those reported in (Moldovan and Novischi, 2004) and (Meyer and Gurevych, 2012) for training-based systems that attempt to disambiguate gloss-based entries in knowledge and lexical resources.

Table 3: Coverage and Precision of each feature individually on the implemented system

	Prec	Cov
SemRel	0.86	0.19
DefSim	0.87	0.29
ShrdLbls	0.84	0.09
MCS	0.75	1.0
SS	0.91	0.21

Next, we examine the effect of the different chosen features on dataset2. We can see in Table 3 the precision and coverage for using each of the features independently on the dataset. SS in the table refers to the instances where only one possible target sense exists. It has a good coverage, when compared to the other compared features, and also gives the best precision. This is expected as for this feature there is only candidate sense in the possible senses list. The MCS feature refers to the most common sense which is selected by our algorithm as the top ordered sense in the possible target senses list. This feature achieves the maximum coverage but obtains the lowest precision. The DefSim feature focuses on the similarities between the glosses of the compared senses. Its obtained precision is relatively high and its coverage is relatively good when compared to the rest of the features. The effectiveness of this feature was anticipated by us when selecting this particular feature as we noticed that a definition of a sense usually is more focused than the context of ambiguous terms that is usually used in the application of Word Sense Disambiguation. The ShrdLbls feature also gives a good precision but it has the lowest coverage.

CONCLUSION

In this article, we have presented a way for automatically aligning the source word senses of the semantic relations within Wiktionary with their suitable target word senses. Effectively, this process leads to the generation of a new type of links that connect the senses of the terms in a semantic relation instead of merely connecting the terms. The method we employed relies on surface-level features from Wiktionary itself without the usage of any external resources or online service. The features we employed include the classification of the semantic relations in Wiktionary, similarity between Wiktionary labels and the senses glosses text. In addition, it also utilizes the categories structure within Wiktionary. We performed an evaluation on two datasets to test the effectiveness of the system we developed and compared it with a baseline and other systems targeting the same problem. The results we obtained illustrate that our method gives similar, or superior results in particular cases, to the other state-of-art systems. The method we describe here in this paper has the advantage of being lightweight in the sense that it uses a small number of features and relies exclusively on Wiktionary without the need of any external resource. This makes it applicable to any language even those which are under-resourced. The obtained results also show that our method can compete and give similar or better results to those of training-based systems targeting the same problem.

ACKNOWLEDGEMENT

This work was sponsored by the Deanship of Scientific Research (DSR), University of Jeddah, Jeddah, Saudi Arabia under grant number (571-968-1436-G). The authors, therefore, acknowledge with thanks DSR technical and financial support.

REFERENCES

Artstein, R. , Poesio, M., 2008. Inter-Coder Agreement for Computational Linguistics. Computational Linguistics **34**, 555–596.

Balamurali, A.R., Joshi, A., Bhattacharyya, P., 2011. Harnessing WordNet Senses for Supervised Sentiment Classification, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1081–1091.

Banerjee, S., Pedersen, T., 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet, in: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing ’02. Springer-Verlag, London, UK, UK, pp. 136–145.

Bouras, C., Tsogkas, V., 2012. A clustering technique for news articles using WordNet. Knowledge-Based Systems **36**, 115–128.

Clark, P., Fellbaum, C., Hobbs, J., 2008. Using and extending WordNet to support question-answering, in: Proceedings of the 4th Global WordNet Conference (GWC’08). Citeseer.

Dandala, B., Mihalcea, R., Bunescu, R., 2013. Multilingual word sense disambiguation using wikipedia, in: Proceedings of the 6th International Joint Conference on Natural Language Processing.

De Belder, J., Moens, M.-F., 2010. Text simplification for children, in: SIGIR: Workshop on Accessible Search Systems 23 July 2010 Geneva, Switzerland.

Gabrilovich, E., Markovitch, S., 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, in: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI'07. USA, pp. 1606–1611.

Garrod, S., Terras, M., 2000. The Contribution of Lexical and Situational Knowledge to Resolving Discourse Roles: Bonding and Resolution. Journal of Memory and Language **42**, 526–544.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explor. Newsl. **11**, 10–18.

Haveliwala, T.H., 2002. Topic-sensitive PageRank, in: Proceedings of the 11th International Conference on World Wide Web, ACM, USA, pp. 517–526.

Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G., 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, Wikipedia and Semi-Structured Resources **194**, 28–61.

Jurafsky, D., Martin, J.H., 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 1st ed. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Komiya, K., Okumura, M., 2012. Automatic Selection of Domain Adaptation Method for WSD using Decision Tree Learning. Journal of Natural Language Processing **19**, 143–166.

Krizhanovsky, A.A., Smirnov, A.V., 2013. An approach to automated construction of a general-purpose lexical ontology based on Wiktionary. J. Comput. Syst. Sci. Int. **52**, 215–225.

Kunze, C., Lemnitzer, L., 2002. GermaNet - representation, visualization, application., in: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002). Spain.

Leffa, V.J., 1992. Making foreign language texts comprehensible for beginners: An experiment with an electronic glossary. System **20**, 63–73.

Li, C.H., Yang, J.C., Park, S.C., 2012. Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet. Expert Systems with Applications **39**, 765–772.

Li, C., Sun, A., Datta, A., 2013. TSDW: Two-stage word sense disambiguation using Wikipedia. J Am Soc Inf Sci Tec **64**, 1203–1223.

Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA.

McCrae, J., Montiel-Ponsoda, E., Cimiano, P., 2012. Integrating WordNet and Wiktionary with lemon, in: Chiarcos, C., Nordhoff, S., Hellmann, S. (Eds.), Linked Data in Linguistics. Springer Berlin Heidelberg, pp. 25–34.

Meyer, C.M., 2013. Wiktionary: The Metalexicographic and the Natural Language Processing Perspective. Technische Universität Darmstadt.

Meyer, C.M., Gurevych, I., 2012. To Exhibit is not to Loiter: A Multilingual, Sense-Disambiguated Wiktionary for Measuring Verb Similarity, Proceedings of COLING 2012, pp. 1763–1780.

Meyer, C.M., Gurevych, I., 2010. Worth Its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet, Gelbukh, A. (Ed.), Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, pp. 38–49.

Miller, G.A., 1995. WordNet: A Lexical Database for English. Commun. ACM **38**, 39–41.

Moldovan, D., Novischi, A., 2004. Word sense disambiguation of WordNet glosses. Computer Speech & Language, Word Sense Disambiguation **18**, 301–317.

Mukherjee, S., Bhattacharyya, P., 2012. WikiSent: Weakly Supervised Sentiment Analysis through Extractive Summarization with Wikipedia, in: Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, pp. 774–793.

Nastase, V., 2008. Topic-driven Multi-document Summarization with Encyclopedic Knowledge and Spreading Activation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08. pp. 763–772.

Navarro, G., 2001. A Guided Tour to Approximate String Matching. ACM Comput. Surv. **33**, 31–88.

Navigli, R., Ponzetto, S.P., 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250.

Niemann, E., Gurevych, I., 2011. The People's Web Meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and Wordnet, in: Proceedings of the Ninth International Conference on Computational Semantics, IWCS '11. USA, pp. 205–214.

Pedersen, T., 2000. A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation, in: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL 2000. pp. 63–69.

Peters, W., Peters, I., 2000. Lexicalized Systematic Polysemy in WordNet, in: Proceedings of the 2nd International Conference on Language Resources and Evaluation.

Pilehvar, M.T., Navigli, R., 2014. A robust approach to aligning heterogeneous lexical resources. Presented at the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014.

Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K., 2001. Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. Information Retrieval **4**, 209–230.

Pustejovsky, J., 2012. Semantics and The Lexicon. Springer Science & Business Media.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Sérasset, G., 2012. Dbnary: Wiktionary as a LMF based Multilingual RDF network, in: Language Resources and Evaluation Conference, LREC 2012.

Sérasset, G., Tchechmedjiev, A., 2014. Dbnary: Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations, in: 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing.

Wang, T., Hirst, G., 2014. Applying a Naive Bayes Similarity Measure to Word Sense Disambiguation, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA

Witten, I., Milne, D., 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links, in: Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy, AAAI Press, USA. pp. 25–30.

Submitted: 14/March/2016
Revised : 30/June/2016
Accepted : 20/Sept/2016