ArabRelat: Arabic Relation Extraction using Distant Supervision

Reham Mohamed, Nagwa M. El-Makky and Khaled Nagi
Department of Computer and Systems Engineering, Alexandria University, Alexandria, Egypt

Keywords: Relation Extraction, Linked Data, DBpedia.

Abstract: Relation Extraction is an important preprocessing task for a number of text mining applications, including: Information Retrieval, Question Answering, Ontology building, among others. In this paper, we propose a novel Arabic relation extraction method that leverages linguistic features of the Arabic language in Web data to infer relations between entities. Due to the lack of labeled Arabic corpora, we adopt the idea of *distant supervision*, where DBpedia, a large database of semantic relations extracted from Wikipedia, is used along with a large unlabeled text corpus to build the training data. We extract the sentences from the unlabeled text corpus, and tag them using the corresponding DBpedia relations. Finally, we build a relation classifier using this data which predicts the relation type of new instances. Our experimental results show that the system reaches 70% for the F-measure in detecting relations.

1 INTRODUCTION

Relation Extraction (RE) is the task of extracting semantic relations between entities from plain text. RE is one of the important tasks in computational linguistics and is considered as a preprocessing task for a number of applications, such as: information retrieval (Waitelonis and Sack, 2012; Hsu et al., 2012), question answering (Unger et al., 2012; Yahya et al., 2012), ontology building (Gupta et al., 2014), etc. Although there are several resources of linked data for different languages, the Arabic resources are still very limited. Therefore, there is a great need for automated methods which extract relations from Arabic text to enrich the Arabic linked data. While several relation extraction systems have been proposed for the English language, Arabic RE systems are still very limited due to the lack of tagged corpora and the challenges of the Arabic language.

Among the challenges of the Arabic language is that **Arabic is highly inflectional and derivational**, which makes its morphological analysis a complex task. Inflectional: where each word consists of a root and zero or more affixes (prefix, infix, suffix). Derivational: where all the Arabic words have root verbs of three or four characters. Also, **Arabic is characterized by diacritical marks (short vowels)**. The same word with different diacritics can express different meanings. Diacritics are usually omitted which greatly increases ambiguity. The **absence of capital letters in Arabic** is an obstacle against accurate named entities recognition. All attempts to make Arabic RE systems rely on small tagged corpora and are limited to a set of relations and specific domains. Examples of these challenges are shown in Figure 1.

In this paper, we introduce *ArabRelat*, an Arabic relation extraction system that adopts the method of distant supervised learning. In distant supervision, a large database of semantic relations is used along with a corpus of unlabeled Web data, such as: Wikipedia. The corpus is used to extract sentences which contain the relation entities. These sentences tagged with the corresponding relation types, are used to build the training data. Finally, the system uses this data to train a Relation classifier which predicts the relation type of new instances. Several features are extracted from the Arabic sentences to build the classifier. Among these features, we extract a set of Arabic-specific rich features which characterize relations in the Arabic language.

We evaluate the system using Arabic DBpedia as the database of semantic relations, and Wikipedia as the untagged corpus. Our results show that the system could achieve 70% F-measure for extracting 97 types of relations which shows its applicability for general relation extraction.

Our main contribution can be summarized in the following points:

1. Building an Arabic relation extraction system using distant supervised learning.

Mohamed, R., El-Makky, N. and Nagi, K..
ArabRelat: Arabic Relation Extraction using Distant Supervision.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 2: KEOD*, pages 410-417
ISBN: 978-989-758-158-8

2. Constructing a relation classifier which predicts the relation type of newly unseen instances.
3. Introducing new Arabic specific features which characterize relations in the Arabic language.

The rest of the paper is organized as follows: Section 2 shows some of the work related to our system. Section 3 shows the Arabic corpora used to construct the training data. Section 4 shows the details of the system architecture. Section 5 shows the different features used to build the relation classifier. In Section 6, we show the results of the system evaluation. Finally, we conclude the paper in Section 7.

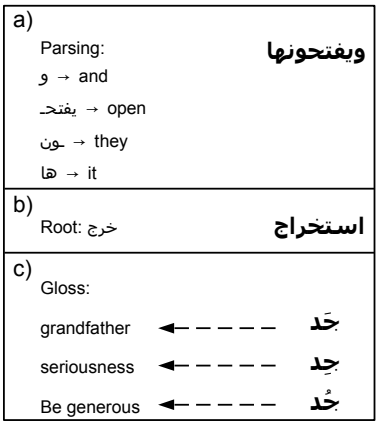

Figure 1: Examples of the challenges of the Arabic language. a) Shows one Arabic word that is chuncked into four English words. b) Shows an example of a three character root of an Arabic word. c) Shows that the diacritics could change the meaning of the same word.

2 RELATED WORK

Some attempts for automatic relation extraction have been proposed in literature. These attempts can be classified into: supervised techniques, unsupervised techniques and distant supervised techniques. (Snow et al., 2004) built a relation classifier that makes a binary decision of whether two nouns are related by hypernym (is-a) relation or not. Given a training set of text containing hypernym pairs, the algorithm automatically extracts useful dependency paths and applies them to new corpora to identify novel pairs. (Banko et al., 2007) built a domain independent system for discovery of relations extracted from text that scales to diversity and size of the web corpus. The system uses self-learning where given a small corpus sample, it outputs a classifier that labels candidate extractions as "trustworthy" or not without hand-tagged data.

(Mintz et al., 2009) introduced the idea of distant supervision. This paper uses Freebase to provide distant supervision for relation extraction where any sentence that contains a pair of entities, that participate in a known Freebase relation, is likely to express that relation in some way. (Nguyen and Moschitti, 2011) proposed a joint model between distant supervised data and manually annotated data from ACE. Their system shows good accuracy for extracting 52 types of relations which suggests the applicability of distant supervision for general RE. In (Fan et al., 2014), the distantly supervised relation extraction was solved as a matrix completion problem. (Yao et al., 2012) uses an unsupervised approach to handle the problem of *Polysemy* where the same pattern can have several meanings. It employs local features and global features to induce pattern senses by clustering feature representations of pattern contexts.

All of the previous systems were built for the English language. Few systems have been proposed for Arabic, however, all of which depend on tagged small corpora. (Alsaif and Markert, 2011) presented an algorithm to identify explicit discourse connectives and the relations they signal for Arabic text. They annotated news articles from Arabic Penn Treebank to build their system. (Kambhatla, 2006) built a minority voting scheme among a committee of classifiers to enhance the recall of the relation classifier. This system was trained and tested using the datasets of ACE 2004 relation extraction task for English, Arabic and Chinese (NIST, 2003).

On the other hand, some systems were proposed to study the linguistic features of the Arabic language and to use these rich features to extract useful information. For example, (Diab et al., 2008) used some Arabic rich morphological features to predict the semantic roles of Arabic text. (Green and Manning, 2010) studied the Arabic linguistic features to achieve better parsing for the Arabic text. To the best of our knowledge, our paper is the first work to exploit the Arabic rich linguistic features to extract relations without using any tagged data.

3 ARABRELAT CORPORA

In this section, we describe the Arabic corpora that we use to build our system.

3.1 Arabic Wikipedia

Wikipedia is one of the most commonly used resources in computational linguistics. The attraction to Wikipedia returns to its large size, its diversity and for being always up to date. The Arabic version of

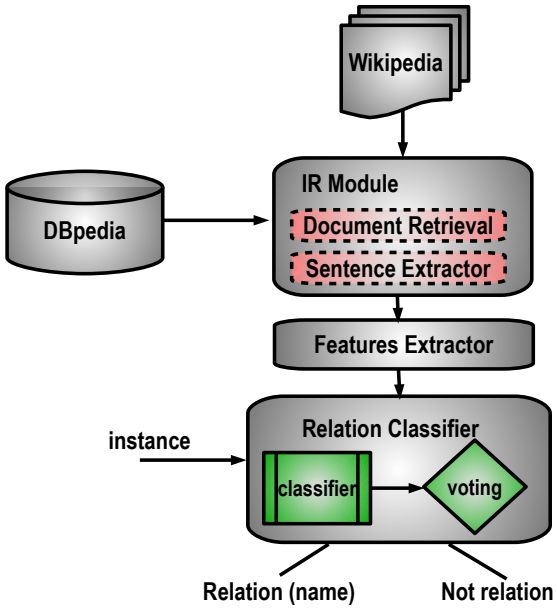

Figure 2: System Architecture.

Wikipedia has over 350,000 articles and is currently the 21^{st} largest edition of Wikipedia[1].

3.2 Arabic DBpedia

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. It is one of the largest resources of linked data. The structured data are extracted from the infoboxes of Wikipedia pages, which are offered as parameter and value. However, the names of these parameters are ambiguous, where the same parameter can be expressed using different names; such as: birthplace and placeofbirth. Therefore, DBpedia Mapping Language has been developed to help in mapping these properties to an ontology. DBpedia is offered in many languages. Arabic DBpedia release is still unavailable on the official DBpedia website, but some unofficial dumps are available.

We use DBpedia as the database of semantic relations, along with Wikipedia as the unlabeled text corpus, to build our distant supervised training data.

4 SYSTEM ARCHITECTURE

The system architecture is shown in Figure 2. The goal is to build a system that extracts new relations from a large text corpora, such as: Wikipedia, such that the relations should be domain independent and the system should not rely on previously tagged data. Therefore, we adopt the idea of distant supervision, where a large database of linked data is used to build a training set using the relations and their entities. For the purpose of this system, we use DBpedia to get the training relations and we extract the corresponding sentences from Wikipedia to build the training set.

The system can be divided into two stages: information retrieval and relation classification. First, the DBpedia relations and Wikipedia pages are fed into an Information Retrieval(IR) module. The main function of the IR module is to retrieve the pages that are semantically related to each relation. We use Wikipedia as an ontology of concepts, to build an inverted index of Wikipedia terms. Then, we convert the relation with its entities into a vector of concepts to retrieve the most relevant pages. The IR module then extracts the sentences that contain the entities of the relation from the retrieved pages. The extracted sentences are tagged according to the relation type. The sentences then pass to a features extraction module which extracts the different features using a morphological analyzer and a dependency parser.

In the second stage, we build a relation classifier which predicts the relation type between two unseen entities. We use the training data constructed in the previous stage to train an SVM classifier, which classifies a Wikipedia sentence containing the two entities into a relation type, after extracting its features. To make the system more robust, we classify all the sentences including the two entities in question. Then, we use a voting scheme which selects the relation class that appears most frequently with the highest confidence. In the next subsections, we explain each module in detail.

4.1 IR Module

As an initial step, we extract the sentences from Wikipedia that correspond to a DBpedia relation. These sentences would be used as the training data tagged with the relation type. This process is referred to as *distant supervision*. Due to the large ambiguity of Arabic language, where one word may refer to several meanings, we do not match a sentence to a relation directly. Instead, we use a semantic information retrieval (IR) module to retrieve the related documents i.e Wikipedia pages that are relevant to the relation. Then we extract the sentences from the retrieved documents. This way we could guarantee the semantic relation between the extracted sentences and the relation.

The details of the IR module is illustrated in Fig-

[1]https://en.wikipedia.org/wiki/Arabic_Wikipedia

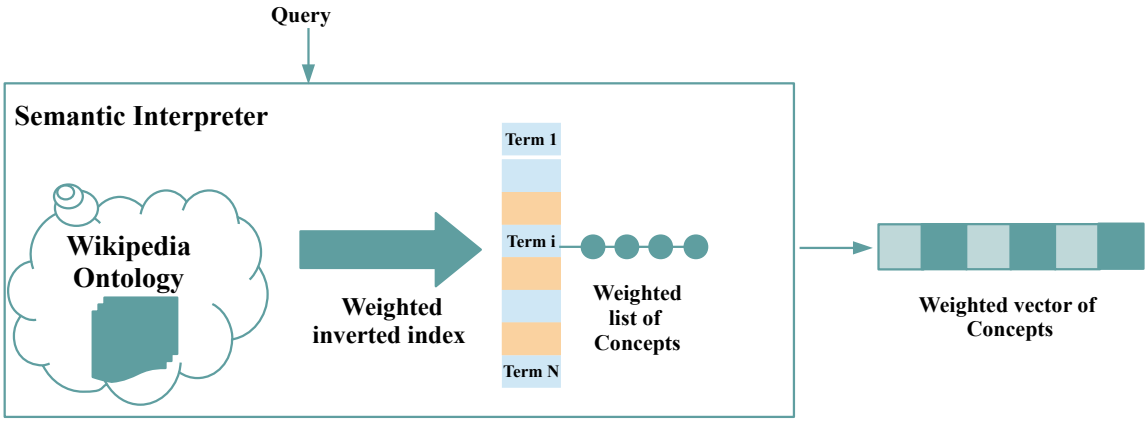

Figure 3: Details of the Information Retrieval (IR) module.

ure 3. We use Wikipedia as an ontology of concepts, where each Wikipedia page represents a concept (Gabrilovich and Markovitch, 2007). Then, we build an inverted index for all Wikipedia terms, such that each term is represented as a vector of concepts. When a query enters the system, it is parsed into terms. Finally, the query is interpreted as one vector using the Cosine similarity of its terms vectors. The concepts (pages) which have the highest weights in the final vectors are the most relevant to the query.

The query here is the DBpedia relation represented by its two entities. For each relation, we use the IR module to get the most relevant Wikipedia pages. Then, we parse each page into sentences using the punctuation marks. The sentences which contain the two entities are used to construct the training data.

4.2 Features Extraction

Several features are extracted from the Arabic sentences, which can be classified into three types: (a) Lexical features, including: part-of-speech tags, types of entities, etc, (b) Syntactic features, including the syntactic path between two entities, sentence voice, etc, and (c) Arabic-specific features which are special features that characterize the Arabic language. More details about the features types are shown in Section 5.

Once the relevant sentences are retrieved from the IR module. Each sentence is converted into a vector of features and tagged with the corresponding relation type. These tagged feature vectors form the final training data which is used to train the Relation classifier.

4.3 Relation Classifier

The goal of the relation classifier is to extract the relation triplets $(entity_1, entity_2, relation\_type)$ from unseen text with high confidence. The constructed training data is used to train an SVM classifier, which classifies one sentence into a relation type. For any two unseen entities, first we extract all the relevant sentences that contain the two entities using the IR module. Then, we classify each sentence using the SVM classifier into a relations type. Finally, we use a voting scheme which selects the most confident relation type of these two entities. The voting scheme calculates the total confidence of each relation type predicted by the SVM classifier and selects the relation type with the highest confidence.

Since we are more concerned about the false positive rate of the overall system, we only detect a relation type between the two entities if the confidence value is larger than a confidence threshold α. Otherwise, the system fails to detect a relation. Although this aggressive way fails to detect all true relations, it guarantees that the detected relations are always true.

5 FEATURES

Table 1 summarizes the features extracted by ArabRelat system.

5.1 Lexical Features

The lexical features describe the two entities and the words around them. This type of features includes:

- Number of words between the two entities.
- Part-of-speech tags of the two entities.
- Named-entity types of the two entities.
- Part-of-speech tags of the words between the two entities.

Table 1: Features extracted by ArabRelat system.

Lexical Features
Number of words between the two entities.
Named-entity type of first entity.
Named-entity type of second entity.
POS tag of first entity.
POS tag of second entity.
POS tags of k words before the first entity.
POS tags of k words after the second entity.
Syntactic Features
Syntactic path between the two entities.
Sentence voice, values are: *passive* or *active*.
isNegated a flag for negated relations.
Arabic-specific Features
Structural word order, values are: *SVO* or *VSO*.
Number matching with the first entity, values are: *singular*, *plural* or *N/A*.
Number matching with the second entity, values are: *singular*, *plural* or *N/A*.
Gender matching with the first entity, values are: *feminine*, *masculine* or *N/A*.
Gender matching with the second entity, values are: *feminine*, *masculine* or *N/A*.
Verb mood, values are: *subjunctive*, *jussive* or *N/A*.

- Part-of-speech tags of k words before the first entity.
- Part-of-speech tags of k words after the second entity.

Where k is a parameter. For our system, we set $k = 3$. These features have been used in previous work, such as: (Mintz et al., 2009), (Yao et al., 2012). The function of the lexical features is to characterize the sentence assuming that sentences with the same relation type would exhibit similar lexical features.

5.2 Syntactic Features

Syntactic parsing is the process of parsing plain text into linguistic units (e.g. words) which are connected using directed links. We use the syntactic path between the two entities, which consists of the sequence of the head words on the directed path between the two entities. To get the syntactic path, we use Stanford Arabic parser (Manning et al., 2014). An example is shown in Figure 4.

We also use the voice of the sentence (active or passive) as a relation feature, where similar relations usually appear in different sentences with the same sentence voice. For example, the relation "was born in" usually appears in passive voice, while the relation "traveled to" usually appears in active voice, even if the entities have the same types in both relations.

We also add a feature that shows whether the relation is negated. If a negative part appears in the words between the two entities, this may indicate that there is no relation between these two entities.

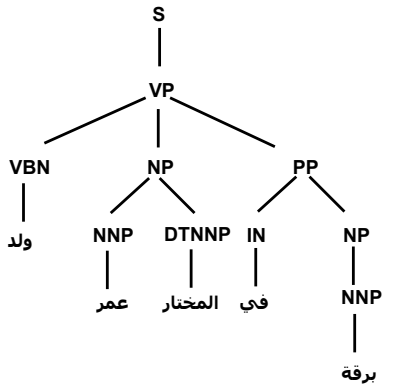

Figure 4: An example of Arabic sentence parsed using Stanford parser.

5.3 Arabic-specific Features

The Arabic language is one of the Semitic languages that have unique characteristics different from the English language. Therefore, we exploit some of the Arabic rich morphological features, which could be used to address the Arabic language challenges and better discriminate entities relation types. Among the Arabic-specific features are:

5.3.1 The Structural Word Order

Arabic sentences differ in the syntactic order of the words. Sentences can be classified into two

types: subject-verb-object (SVO) sentences and verb-subject-object (VSO) sentences. Relations of the same type usually appear in the same form. Therefore, we use the type of the sentence as a feature in our relation classifier.

5.3.2 Number Matching

The Arabic verbs include number information. For example, the verb differs based on the subject number (singular or plural). Therefore, among the rich features that could characterize a relation, is the number matching between the relation verb (i.e. the verb between the two entities) and the two entities. Using this property, we extract two more features:

- Number matching between the verb of the relation (if any) and the first entity.
- Number matching between the verb of the relation (if any) and the second entity.

5.3.3 Gender Matching

The Arabic verbs are also characterized by the gender information, where the first letter of the verb differs according the gender of its subject. As most of the relations are characterized by a verb describing this relation, we argue that the matching pattern between the verb gender and the gender of the entities could indicate the relation between these entities. Therefore, we add another two features:

- Gender matching between the verb of the relation (if any) and the first entity.
- Gender matching between the verb of the relation (if any) and the second entity.

5.3.4 Verb Mood

The Arabic verbs also differ according to the verb mood. Arabic verb moods include: subjunctive and jussive. Relation verbs usually come in one form according to the relation type. Therefore, the verb mood is also added as a discriminative relation feature.

6 PERFORMANCE EVALUATION

6.1 Implementation

We use Apache Lucene library [2] for the IR module. We use Lucene indexer to build the inverted index over Wikipedia pages and the searcher to get the most relevant pages to a relation query. For building the classifier, we use the SVM implemenation of Weka library (Hall et al., 2009).

For the features extraction, we use Stanford parser (Manning et al., 2014) to extract the syntactic path. The morphological analyzer, MADAMIRA (Pasha et al., 2014), is used to extract all the other features. MADAMIRA provides a large set of rich morphological features of the Arabic text, including: stem, root, lemma, POS tags, gloss, case, mood, etc. The last version of MADAMIRA also provides named-entity tagging. So we use it to extract the morphological features and named entities.

6.2 Datasets

We build the training data using DBpedia relations and sentences extracted from Wikipedia. We use a subset of DBpedia dump consisting of 1358 relation instances which correspond to 97 different relation classes. We divided the instances of each relation type equally into training and testing. For each relation instance, we extracted all the relevant sentences from Wikipedia. The total number of sentences of the training data is 4915. The total number of sentences of the test data is 7500.

We also extracted some negative relations, which are unrelated entity pairs that exist in one sentence. To build the negative relations, we used the entities that appear in one Wikipedia sentence and do not appear in the whole DBpedia relations. Although one might criticize this method since DBpedia is incomplete, which means that the negative relations may in fact express a relation, we argue that this will lead to a decrease in the true positive rate, while maintaining a low false positive rate, which is our main concern in the system.

6.3 Evaluation Results

Two evaluation methodologies are used to evaluate the system. In the first method, half the instances of each relation are used in training and the other half is held out for testing. In this method, we trust the automatic tagging manipulated by our system, thus we call it *Trust method*. In the second method, we use human evaluation where a small subset of the test relations are tagged by an Arabic speaker and used to evaluate the system. We call this method *Human evaluation method*.

6.3.1 Trust Method

We compare ArabRelat system against a baseline relation classifier. The baseline uses ArabRelat system

[2] https://lucene.apache.org/

Table 2: Results of the test dataset.

	Baseline	**ArabRelat**
Precision	0.59	0.74
Recall	0.21	0.67
F-measure	0.31	0.70

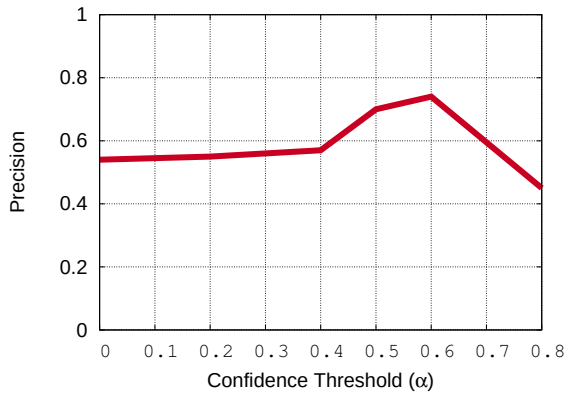

Figure 5: Effect of confidence threshold α on the system accuracy.

with subset of the features. We assume the baseline features are the lexical and syntactic features. We show the effect of adding the Arabic-specific features on the system accuracy. Table 2 shows the results of the system using the evaluation test set. The results show that while the baseline maintains good precision, it bitterly decreases the recall. ArabRelat improves the precision by 15% over the baseline, and improves the recall by 46%.

6.3.2 Effect of Confidence Threshold

Figure 5 shows the effect of the confidence threshold α on the system accuracy using the test data. As the value of α increases, the accuracy of the system increases until it reaches its optimal value at $\alpha = 0.6$. For larger values of alpha the accuracy decreases because the number of instances which survive becomes very small, thus more prone to false positive errors. We set the default value of α to 0.6.

6.3.3 Human Evaluation Method

Due to the lack of public gold-standard Arabic relation data, we construct another test dataset tagged by an Arabic speaker. We extracted 100 sentences of the test dataset, an Arabic native speaker tagged each sentence to a relation type. The speaker was given each sentence and the two entities to be tagged, with a set of relation types. The task was to tag each sentence with a suitable relation type or none if the sentence does not express a relation between the two entities.

Table 3: Human Evaluation Results.

	Baseline	**ArabRelat**
Precision	0.34	0.50
Recall	0.33	0.43
F-measure	0.34	0.46

We used a subset of 18 relation types. The results are shown in Table 3. The precision of ArabRelat system decreases due to the small size of the dataset, however, it still outperforms the baseline.

7 CONCLUSION

In this paper, we propose a novel Relation Extraction system for the Arabic language. The system uses distant supervised learning to build a relation classifier, without the need of prior labeled data. We introduce new Arabic specific features that characterize Arabic relations. Our experimental results on sentences extracted from Wikipedia show that the system achieves 70% overall F-measure for detecting 97 relation types.

REFERENCES

Alsaif, A. and Markert, K. (2011). Modelling discourse relations for arabic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction for the web. In *IJCAI*.

Diab, M. T., Moschitti, A., and Pighin, D. (2008). Semantic role labeling systems for arabic using kernel methods. In *ACL*.

Fan, M., Zhao, D., Zhou, Q., Liu, Z., Zheng, T. F., and Chang, E. Y. (2014). Distant supervision for relation extraction with matrix completion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*.

Green, S. and Manning, C. D. (2010). Better arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 394–402. Association for Computational Linguistics.

Gupta, R., Halevy, A., Wang, X., Whang, S. E., and Wu, F. (2014). Biperpedia: An ontology for search applications. *Proceedings of the VLDB Endowment*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data min-

ing software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hsu, I.-C., Lin, H.-Y., Yang, L. J., and Huang, D.-C. (2012). Using linked data for intelligent information retrieval. In *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on*. IEEE.

Kambhatla, N. (2006). Minority vote: at-least-n voting improves recall for extracting relations. In *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*. Association for Computational Linguistics.

Nguyen, T.-V. T. and Moschitti, A. (2011). Joint distant and direct supervision for relation extraction. In *IJCNLP*.

NIST, U. (2003). The ace 2003 evaluation plan. *US National Institute for Standards and Technology (NIST)*.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. M. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*.

Snow, R., Jurafsky, D., and Ng, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A.-C., Gerber, D., and Cimiano, P. (2012). Template-based question answering over rdf data. In *Proceedings of the 21st international conference on World Wide Web*, pages 639–648. ACM.

Waitelonis, J. and Sack, H. (2012). Towards exploratory video search using linked data. *Multimedia Tools and Applications*.

Yahya, M., Berberich, K., Elbassuoni, S., Ramanath, M., Tresp, V., and Weikum, G. (2012). Natural language questions for the web of data. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics.

Yao, L., Riedel, S., and McCallum, A. (2012). Unsupervised relation discovery with sense disambiguation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics.