Confidence Measure Estimation for Open Information Extraction

Vahideh Reshadat*
Malek-Ashtar University of Technology, Tehran, Iran
vreshadat@mut.ac.ir
Maryam Hoorali
Malek-Ashtar University of Technology, Tehran, Iran
mhourali@mut.ac.ir
Heshaam Faili
School of Electrical and Computer Engineering, College of Engineering,, University of Tehran, Tehran, Iran
hfailii@ut.ac.ir

Received: 22/Jul/2017 Revised: 16/Sep/2017 Accepted: 24/Dec/2017

Abstract

The prior relation extraction approaches were relation-specific and supervised, yielding new instances of relations known a priori. While effective, this model is not applicable in case when the number of relations is high or where the relations are not known a priori. Open Information Extraction (OIE) is a relation-independent extraction paradigm designed to extract relations directly from massive and heterogeneous corpora such as Web. One of the main challenges for an Open IE system is estimating the probability that its extracted relation is correct. A confidence measure shows that how an extracted relation is a correct instance of a relation among entities. This paper proposes a new method of confidence estimation for OIE called Relation Confidence Estimator for Open Information Extraction (RCE-OIE). It investigates the incorporation of some proposed features in assigning confidence metric using logistic regression. These features consider diverse lexical, syntactic and semantic knowledge and also some extraction properties such as number of distinct documents from which extractions are drawn, number of relation arguments and their types. We implemented proposed confidence measure on the Open IE systems' extractions and examined how it affects the performance of results. Evaluations show that incorporation of designed features is promising and the accuracy of our method is higher than the base methods while keeping almost the same performance as them. We also demonstrate how semantic information such as coherence measures can be used in feature-based confidence estimation of Open Relation Extraction (ORE) to further improve the performance.

Keywords: Information Extraction; Open Information Extraction; Relation Extraction; Knowledge Discovery; Fact Extraction.

1. Introduction

Information Extraction is the task of automatically extracting structured data from unstructured text. One of the core information extraction tasks is relation extraction which aims at extracting semantic relations among entities from natural language text. Relation extraction can potentially benefit a wide range of NLP tasks such as: Web search, question answering, ontology learning, summarization, building knowledge bases, etc. [1,2].

The huge and fast-growing scale, a mixed genre of documents and infinite types of relations are challenges of the Web-scale relation extraction [3]. The traditional approaches to information extraction (such as [4-6]) assume a fixed set of predefined target relations and usually don't scale to corpora where the number of target relations is very large [7,8]. An alternative paradigm, Open Information Extraction (OIE) aims to scale information extraction methods to the size and diversity of the Web corpus. Open IE systems extract relational tuples from text, without requiring a pre-specified vocabulary [9-12].

The key goals of Open IE are: (1) domain independence, (2) unsupervised extraction, and (3) scalability to large amounts of text [13]. Since Open IE is never perfectly accurate, it is helpful to have an effective measure of confidence.

Following [14], there are at least three important applications of accurate confidence estimation. First, accuracy-coverage trade-offs are a common way to improve data integrity in databases. Efficiently making these trade-offs requires an accurate prediction of correctness. Second, confidence estimates are essential for interactive information extraction, in which users may correct incorrectly extracted fields. These corrections are then automatically propagated in order to correct other mistakes in the same record. Directing the user to the least confident field allows the system to improve its performance with a minimum amount of user effort. Third, confidence estimates can improve performance of data mining algorithms that depend upon databases created by information extraction systems [15]. Confidence estimates, provide data mining applications with a richer set of "bottom-up" hypotheses, resulting in more accurate inferences.

* Corresponding Author

This paper focuses on the confidence estimation for Open IE systems. In this work we use logistic regression, a probabilistic machine learning model, to automatically assign a confidence weight to an extraction. This paper makes the following contributions:

- This paper proposes several diverse new lexical, syntactic and semantic features for estimating confidence of open relation extraction systems using a probabilistic model.
- We study how the proposed features for weighting extracted relations affect the performance of results and use a logistic regression classifier to assign a confidence score to each Open IE extraction in order to improve precision.
- Our evaluations show that the proposed method can drop noisy extractions from Open IE systems' outputs and demonstrate that effective incorporation of diverse features enables our approach to identify correct instances with more certainty.

The rest of this paper is organized as follows. Section 2 presents related work. Proposed methodology is described in Section 3. We present results of our experiments in Section 4 and end with conclusion and future work in Section 5.

2. Related Works

In this section we review some open information extraction systems with respect to confidence estimation.

WOE_{pos} [16] applies Wikipedia for self-supervised learning of unlexicalized extractor and is limited to light features such as Part-Of-Speech (POS) tags. WOE_{pos} generates relation-specific training examples by matching Infobox attribute values to corresponding sentences and abstracts these examples to relation-independent training data to learn an unlexicalized extractor. WOE_{parse} [16] is a pattern classifier learned from dependency path patterns which uses typed dependencies as features. Authors in their evaluation showed that using deep syntactic parsing improves the precision of their system, however at a high cost in extraction speed.

R2A2 [17] exploits an argument learning component. It makes use of a number of classifiers to identify the arguments of a verb phrase (based on hand-labeled training data). Two classifiers identify the left and right bounds for first argument and one classifier identifies the right bound of second argument.

ZORE [18] is a syntax-based Chinese open relation extraction system for extracting semantic patterns and relations from Chinese text. ZORE realizes relation samples from automatically parsed dependency trees, and then extracts relations with their semantic patterns iteratively through a double propagation algorithm. [19] also considers Chinese Open relation extraction. It can be assumed as a pipeline of word segmentation, POS and parsing.

LSOE [20] is an Open IE extractor based on lexical-syntactic patterns. It provides a plain solution to perform rule-based extraction of facts using POS-tagged text. The method was developed based on two types of patterns: (1) generic patterns (2) rules from Cimiano and Wenderoth proposal [21]. LSOE performance was compared with ReVerb and DepOE. The results show that LSOE extracts relations that are not learned by other extractors and achieves compatible precision.

Wanderlust [22] uses hand-labeled training data to learn extraction patterns on the dependency tree. After annotating 10,000 sentences parsed with LinkGrammar, it learns 46 general linkpaths as patterns for relation extractions.

Some Open IE methods are designed to obtain binary facts and they usually don't capture higher order N-ary facts. KrakeN [23] considers this weakness. It can extract more facts per sentence in high precision and is capable of extracting unary, binary and higher order N-ary facts. Since using a dependency parser results in cost in recall and speed, many sentences were ignored due to heuristic of detecting erroneous parses. OLLIE [9] aims to improve the Open IE systems by using a hybrid approach based on bootstrapping. It learns pattern templates automatically from a training set that is bootstrapped from relations extracted by the ReVerb system. It obtains the pattern templates from the dependency path connecting pairs of entities and their corresponding relations. The patterns are then applied over the corpus and new facts are obtained.

ClauseIE [13] is a novel, clause-based approach to open information extraction which differs from previous approaches in that it separates the detection of "useful" pieces of information expressed in the sentence from their representation in terms of extractions. ClauseIE exploits linguistic knowledge about the grammar of the English language to first detect clauses in an input sentences and to subsequently identify the type of each clause according to the grammatical function of its constituents. ClauseIE attains high precision and recall and can be customized to output triples or n-ary facts. EXEMPLAR [24] is an ORE approach that extracts n-ary relations. It uses rules over dependency parse trees to detect relation instances. EXEMPLAR's rules are used to each candidate argument separately as opposed to all candidate arguments of an instance. Since the aim is to gain low computational cost and high precision, its variations have been indicated by different dependency parsers. The results are promising and EXEMPLAR outperforms the systems that support n-ary extraction.

Bast and Haussman [25] proposed a method called CSD-IE that uses contextual sentence decomposition for Open IE. It decomposes a sentence into the parts that semantically 'belong together'. The facts are then captured by recognizing the (implicit or explicit) verb in each part. In [26], the same authors improved the informativeness of extracted facts in Open IE by using some inference rules. Uninformative extracted facts are obstacle for semantic search applications utilizing them. Their evaluation shows that this approach can increase the number of correct and informative triples by 15% discarding the uninformative ones [27].

In [28] authors proposed an Open IE system based on semantic role labeling (SRL). They constructed novel extractors based on two semantic role labeling systems, one developed at UIUC's publicly available SRL system [29] and the other at LUND [30].

Existing Open Information Extraction systems have mainly focused on Web's heterogeneity rather than the Web's informality. The performance of the ReVerb system, drops dramatically as informality increases in Web documents. In [31] a Hybrid Ripple-Down Rules based Open Information Extraction (Hybrid RDROIE) system was proposed, which uses RDR on top of a conventional Open IE system. The Hybrid RDROIE system applies RDR's incremental learning technique as an add-on to the state-of-the-art ReVerb Open IE system to correct the performance degradation of ReVerb due to the Web's informality in a domain of interest. The Hybrid RDROIE system doubled ReVerb's performance in a domain of interest after two hours training.

We proposed two preliminary models called TR-DOE and RV-DOE [12]. These two kinds of hybrid systems are made of two shallow and deep Open IE systems by using two combination parameters separately. We detected the best trade-off between precision and recall. Experiments indicate that the proposed hybrid methods obtain significantly higher performance than their constituent systems. The best result was for TR-DOE which had an F-measure almost twice that of TextRunner.

Dependency-based Open information Extraction (DepOE) [32] is a multilingual OIE system based on fast dependency parsing which has the main feature of being able to operate at Web-scale. It uses DepPattern [33], a multilingual dependency-based parser, to analyse sentences and obtain fine-grained information. Then, a small set of extraction rules is applied and the target verb-based triples are generated. There is a more recent version of DepOE system, called ArgOE [11]. ArgOE is a multilingual rule-based OIE method that obtains as input dependency parses in the CoNLL-X format, recognizes argument structures within the dependency parses, and extracts a set of basic propositions from each argument structure. This method does not need training data and has higher recall and precision than previous approaches relying on training data.

Estimating a confidence score for Open Information systems is not addressed in literature so well. TextRunner [34], is first and high scalable Open IE system where the facts are assigned a probability. It counts the number of distinct sentences from which each extraction was found. Assessor uses these counts to assign a probability to each tuple using the probabilistic model.

ReVerb [27] is a strong and successful shallow Open IE system. It makes use of a simple POS tag sequence as a syntactic constraint in order to extract relation phrases and eliminate incoherent extractions and also reduce uninformative extractions. ReVerb uses a classifier to determine a confidence score for each triple. It employs a set of relation independent features and a training set containing 1,000 sentences from the Web and Wikipedia to assign a confidence score to each extraction.

OLLIE [9] is an Open IE system that learns pattern templates automatically from a training set that is bootstrapped from relations extracted by the ReVerb system. It uses a supervised classifier for confidence function. The classifier applies a set of lexical features such as frequency of the extraction patterns, position of function words etc.

Some related works to open relation extraction systems are semantic best-effort information extraction approaches. KnowItAll is a Web extraction system which labels its own training data. It aims to automate and simplify the process of extracting large scale relations from the Web. Its hypothesis is that extractions drawn more frequently from distinct sentences in a corpus are more likely to be correct.

In [14] authors showed conditional random field is an empirically sound confidence estimator for finite state information extraction systems. It has an average precision of 97.6% for estimation field correction. Scheffer et al. describes a confidence estimation algorithm using hidden Markov models in information systems in [35]. They estimate the confidence of only singleton tokens by the difference between the probabilities of their first and second most likely labels.

URNS [36] is a combinational "balls-and-runs" model that evaluates the impact of redundancy, sample size and corroboration from several distinct extraction rules on the confidence score. It was illustrated experimentally that the model's log likelihoods for unsupervised information extraction are considerably higher than previous methods.

In [37] Agichtein proposed an expectation-maximization algorithm for automatically assessing the quality of the extraction patterns and relation tuples for partially supervised relation extraction. This method was evaluated for different types of patterns and improved extraction accuracy over heuristic-based methods.

3. Proposed Method

In this section we describe our proposed approach for assigning probability of correctness to Open IE systems' extractions.

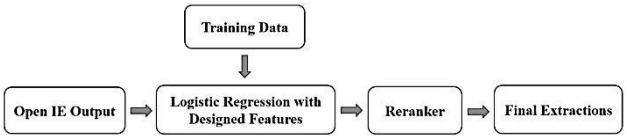

Fig. 1. The outline of Relation Confidence Estimator for Open Information Extraction (RCE-OIE)

There are various parameters that can aid in detecting accurate relations. This idea inspires us to develop a learning-based approach that applies our proposed parameters as features to assign a weight of correctness to the extracted semantic relations. Based on this assumption, the extractions with high precision will be obtained. The outline of the proposed method is shown in Figure 1.

Our proposed approach takes as input an output of Open IE system and trains a confidence metric on a labeled data set and uses the classifier's weight to assign a confidence score to each extraction.

3.1 Relation Confidence Estimator

Logistic regression is a conditional probability model which is used as the confidence classifier and is the main part of our approach. Relation Confidence Estimator for Open Information Extraction (RCE-OIE) reads every relation instance sequentially. For each relation instance, the confidence classifier computes the probability of its correctness.

This approach focuses on the confidence estimation for output instances of Open IE systems. We use logistic regression, a probabilistic machine learning model, to automatically assign a score to each input relation instance. Logistic regression belongs to the family of classifiers known as the exponential or log-linear classifiers. Like naive Bayes, it works by extracting some set of weighted features from the input, taking logs, and combining them linearly. It classifies an observation into one of two classes. In order to train a model to classify with minimum error as possible, the cost function should be minimized. Gradient descent is our learning algorithm that finds values for the parameters that result in best parameter values and a smaller minimum error. For this purpose we used Weka for implementation.

We formulate the relation confidence estimation for OIE systems as a classification problem by logistic regression classifier. Given the features and weights, our goal is to choose a class (confident or unconfident) for the relation instance. The probability of a particular class given the observation x is:

$$F(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

The logistic function maps values from -∞ and ∞ to lie between 0 and 1.

The problem addressed by the classifier is selecting proper class for each input relation instance in order to maximize the number of correctly extracted instances and thus effectiveness. This model can take our proposed features and return the probability that a particular observation is true and should be considered as a correct instance. We detect the best trade-off between effectiveness and efficiency (computational cost).

Deep features could improve precision and recall over shallow syntactic features, but at the cost of speed. For instance, parser-based features can help to recognize complicated and long distance relations in difficult sentences. Such cases usually cannot be detected by shallow features. Regarding the computational cost associated with rich syntactic features, we used about 14 light-weight features. All features are scalable, domain independent and can be evaluated at extraction time without use of expensive tools. These features could be extracted from the underlying systems.

3.2 Proposed Features

We designed various lexical, syntactic and semantic features to the classifier. All features are scalable, domain independent and can be evaluated at extraction time without use of expensive tools. These features are described in the following.

- Document frequency (D_f): This feature is based on the intuition that a valid relation phrase is found repeatedly in different documents in huge corpora such as the Web. More particularly, this feature considers redundancy impact on the probability of correctness and is defined as the number of distinct documents from which each extraction is found relative to the total number of documents.

$$Doc_f = \frac{|D_r|}{|D|} \quad (2)$$

$|D|$ is the total number of documents and $|D_r|$ is the number of documents containing relation r.

- Type frequency (T_f): This feature accounts for the number of domains in which the relation appears. We used Stanford NER for assigning types for arguments. It assigns one of the seven types (Location, Person, Organization, Money, Percent, Date, Time) to each argument. The arguments which are not in these classes are assigned "Other" tag. Let T={Location, Person, Organization, Money, Percent, Date, Time, Other}. Let domain type (DT) be the set of all possible relations' domain types (DT= T ×T). We use the frequency of domain types of arguments which reveals in the context of a relation. This feature is denoted as T_F and is defined as:

$$T_f = \frac{|AT_R|}{|DT|} \quad (3)$$

Where $|AT_R|$ is the number of distinct domain types that a relation takes and $|AT|$ is the total number of domain types.

Domain Entity frequency (DEf): This feature also considers the types of relation arguments and counts the number of distinct entity pairs of the type of relation's arguments which appear with it. This intuition is similar to that offered by Mesquita [38] to assign weights to the terms in the context of an entity pair in clustering task which could achieve high performance. In this case, we consider relation instead of term and define it as:

$$DE_f = \frac{|R_{at}|}{|R|} \quad (4)$$

|R| is the frequency with which a relation r appears in the context of the arguments of any domain type. It shows the total number of occurrences of a relation with distinct arguments. $|R_{at}|$ is the frequency with which relation r appears in the context of the arguments of its current domain type. It is the total number of occurrences of a relation with distinct arguments of the type of its current arguments.

- Arguments frequency (A_f): This feature is based on the number of distinct arguments that a relation takes and is defined as:

$$A_f = \frac{|A_r|}{|A|} \quad (5)$$

|A| is the total number of all distinct arguments in corpus and $|A_r|$ is the count of distinct arguments which a relation takes.

- Arguments' Coherence: Coherence measures can be applied to automatically rate quality of topics computed by topic models [39]. A set of statements or facts is said to be coherent, if they support each other. We use the C_v and C_a measures proposed by Röder and his colleagues [39]. The framework of these coherence measures is a composition of four parts which differs in segmentation and probability calculation of words. One of the advantages of these measures is that they are based on word co-occurrence statistics estimated on Wikipedia and can detect coherence of proper nouns. Given two argument word sets, we calculate the coherence of each word in the first argument's word set with the words in the second argument's word set mutually and then compute the average of it. It measures the degree that two arguments are supported by each other.

We considered some syntactic and sentence-based features which are described in the follow.

- Arguments and relation covers all words in the sentence.
- There is a verb after second argument in the sentence.
- There is a preposition *in* or *to* after second argument in the sentence.
- First argument contains pronoun.
- There is a *in* or *if* before the first argument in the sentence.
- There is a *that* pronoun before the relation in the sentence.
- There is a *that* or *to* after the second argument in the sentence.
- Number of words in the sentence is less than ten.

We study how using these features for weighting extracted relations affects the precision of results. The next section gives more details about the results of our experiments.

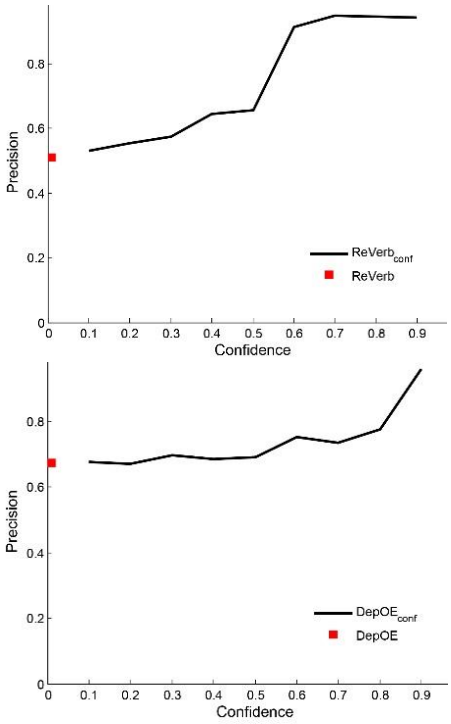

Fig. 2. ReVerb, DepOE and TextRunner have higher precision than their base systems for all thresholds.

4. Experiments and Results

In this section, we first describe benchmark datasets and performance metrics, and then give the results obtained by our approach and its counterparts.

4.1 Dataset

We used the dataset that was provided by Fader and his colleagues [27] in our experiments. They created a test set of 500 sentences sampled from the Web, using Yahoo's random link service. This dataset contains the output of the different extractors run (such as TextRunner and ReVerb) on the 500 selected sentences. Two human judges independently evaluated each extraction as 'correct' or 'incorrect'. The judges reached agreement on 86% of the extractions, with an agreement score of κ=0.68. The subset of the data where the two judges concur, is used in our experiments. The judges labeled uninformative extractions (where critical information was dropped from the extraction) as incorrect. This is a stricter standard than was used in previous Open IE evaluations [17].

In this collection, the extractions from a set of 1000 sentences from the Web and Wikipedia are available. The classifier was trained on 1000 random Web sentences with the proposed features. To collect syntactic features, we need to perform POS tagging and chunking therefore we use OpenNLP package [1]. Since the dataset only contains sentences, document frequency is estimated by assuming each sentence as a document.

1. http://opennlp.sourceforge.net

4.2 Performance Measures

In the experiments, we conducted evaluations using two important criteria: precision and F-measure. For more detail about these metrics refer to [40]. The quality of the results is evaluated by comparing the relation instance pairs obtained by the system and those in the ground truth annotated by annotators. Formally, precision (*P)* and F-measure is defined as follows:

$$P = \frac{|S \cap G|}{|S|} \quad (6)$$

$$F - measure = 2 \times \frac{P \times R}{P + R} \quad (7)$$

Where *S* is the set of relation instances generated by the system, and *G* is the set of correct labeled relation instances in the annotated gold standard set. R denotes recall, which is the ratio of the number of correct extractions retrieved to the total number of correct extractions in the dataset.

4.3 Experiment Results and Discussion

We evaluate the effect of applying logistic regression classifier, a linear regression method, on the output of different Open IE systems with the aid of some features and explore the behavior of it. We compare performance of TextRunner, ReVerb and DepOE with their confidence-based status.

A confidence score is assigned to each extraction using a classifier trained on mentioned training set with proposed features. Figure 2 reports the detailed precision curves of some Open IE systems with different confidence thresholds. Precision is the ratio of the number of correct extractions retrieved to the total number of extractions retrieved. The system names with *conf* subscript focuses on using only extractions with confidence values equal or above a threshold and ignores other extractions. As these figures show, precision curves always have higher levels of precision than their base for all confidence thresholds and all systems. This shows the effectiveness of proposed confidence score. Actually, the proposed method focuses on increasing the precision and uses the confidence as a filter policy to decrease the number of incorrect extractions and increase the correct ones, as a result, leads to the high precision.

DepOE's base system has higher level of precision than those of ReVerb and TextRunner. This is mainly because parser-based features used by DepOE are useful for handling correct extractions and thus, overall precision of it is high.

ReVerb and TextRunner start at low precision due to intrinsic weakness of shallow extractors in detecting relation instances.

Variations of precision for different values of confidence thresholds are also shown for all systems in Figure 2. When the confidence threshold is low, most of the extractions are considered as confident and the amount of precision for all systems is near the precision of their base cases. As confidence grows, the number of included extractions is gradually decreased but most of them are regarded correct therefore the precision slowly increases as confidence threshold increases. Figure 2 also shows that precision increases as confidence threshold increases but the slope of TextRunner and ReVerb's precision curves increase quicker than that of DepOE. Due to deep features used in DepOE, it extracts accurate triples and initial precision of it is higher than others and has relatively high start point difference with other approaches. It starts at high precision due to discarding of potentially low quality extractions from it. Thus, the proposed approach improves performance of both shallow and deep extractors. When confidence increases, the precision curves also increase as a result of filtering incorrect extractions.

The value of the threshold was examined from 0.1 to 0.9 by increments of 0.1. We examined F-measure values for all thresholds and found the maximum amount of it for each method. The results were shown in Figure 3. The F-measure is the uniformly weighted harmonic mean of the precision and the recall. Determination of the maximum value for F-measure is an attempt to find the best possible trade-off between recall and precision.

All systems achieve almost the same F-measure as their base. It shows that, the proposed method can achieve reasonable F-measure, but with more confidant extractions. Because of the deep tools used in the structure of deep_extractors, DepOE has the best F-measure in comparison with the other systems.

DepOE produces a little bit lower F-measures in comparison with its base case. The proposed method provides a boost in F-measures of the shallow extractors. TextRunner and ReVerb achieve an F-measure that is slightly higher than their base cases. This is mainly because of the depth of tools applied in their structures.

Features and learned weights in the logistic regression classifier are shown in Table 1. All of these features are effectively calculable and derived from corpus and sentences structures.

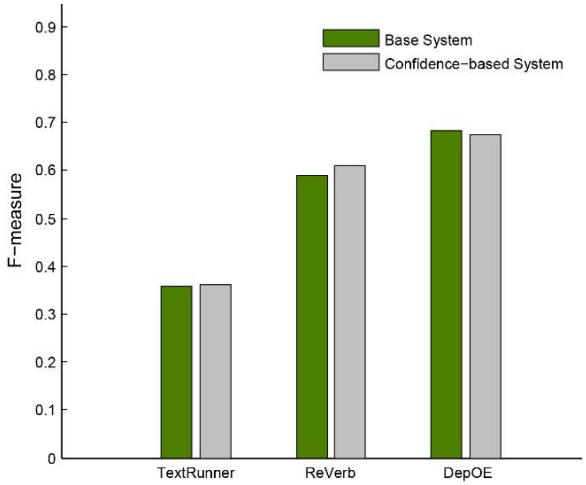

Fig. 3. All systems achieved approximately the same levels of F-measure

Since ReVerb as a robust and successful Open IE system is the nearest related work to our approach, we compare our method with it. ReVerb was applied on the test set of 500 sentences and the resulting extractions were used. We used Reverb's and our proposed confidence score and examined different threshold values to assess the precision variations. Our preliminary results from an analysis of ReVerb's output are reported in Figure 4.

The number of extractions with high confidence decreases as confidence threshold values increase and also the number of correct extractions increases as far as about all of retrieved

Table 1. Confidence classifier assigns a confidence score to an extraction from a sentence using these features.

Weight	Feature
0.01	D_f
0.12	T_f
0.24	DE_f
0.32	A_f
0.5	Cv
0.41	Ca
0.5	Arguments and relation covers all words in the sentence
0.49	There is a verb after second argument in the sentence.
-0.56	There is a preposition *in* or *to* after second argument in the sentence.
0.14	First argument contains pronoun.
-0.43	There is a *in* or *if* before the first argument in the sentence.
-0.61	There is a *that* pronoun before the relation in the sentence.
-0.42	There is a *that* or *to* after second argument in the sentence.
1.12	Number of words in the sentence is less than ten.

extractions are correct in high confidence thresholds. Except for extractions with a confidence values near 0.3, the precision of $ReVerb_{PROPconf}$ is always higher (or equal) than that of $ReVerb_{RVconf}$. This shows the effectiveness of proposed features. It seems that by increasing the number of effective features and the size of training data set, the results improves.

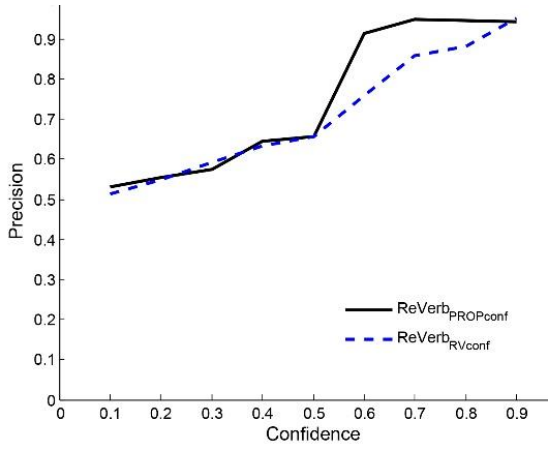

Fig. 4. Precision variation over different confidence values for $ReVerb_{PROPconf}$ and $ReVerb_{RVconf}$

4.4 Evaluating Classifiers

We modeled relation confident for OIE systems' outputs. Our modeling of relation confident was binary: relations are confident or unconfident. Given a corpus, proposed approach should select confident relations to maximize the number of correctly extracted instances.

Table 2 shows the distribution of the values in the confusion matrix for the confidence classifier. The results show that the dominant error in the classifier is classifying an unconfident extraction as confident.

Table 2. The confusion matrix for the performance of the confidence classifier

Gold/Classified	Confident	Unconfident
Confident	65.7%	23.9%
Unconfident	33.1%	77.3%

5. Conclusion

All of Open IE systems make errors and one of the important problems for an Open IE system is specifying the probability that extracted information is correct. In this paper, we used a logistic regression classifier to provide a confidence score for each relation of Open Information Extraction systems where diverse features are employed. It covers a wide range of features from syntactic and sematic (e.g., arguments' coherence) to sentence and corpus based ones (e.g. number of relation arguments and their type). Our evaluations show that effective incorporation of diverse features enables our approach outperform the base Open IE systems in terms of performance. Moreover, proposed features produce results comparable to the confidence score of ReVerb.

We plan to explore utilizing some more efficient features to improve performance of learned model. Furthermore, we are interested to extend experiments to other open IE systems and apply our model to their extractions. We also need to take into consideration the impact of training data set size and do experiments with larger amounts of training data to see if our new implementation improves. Another direction for improvement is to expand the type space of arguments with resources of semantic knowledge such as ontologies.

References

[1] L. Yao, A. Haghighi, S. Riedel, and A. McCallum, "Structured relation discovery using generative models," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1456-1466.

[2] J. Piskorski and R. Yangarber, "Information extraction: Past, present and future," in Multi-source, Multilingual Information Extraction and Summarization, ed: Springer, 2013, pp. 23-49.

[3] B. Min, S. Shi, R. Grishman, and C.-Y. Lin, "Towards Large-Scale Unsupervised Relation Extraction from the Web," International Journal on Semantic Web and Information Systems (IJSWIS), vol. 8, pp. 1-23, 2012.

[4] R. C. Bunescu and R. J. Mooney, "A shortest path dependency kernel for relation extraction," in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 724-731.
[5] A. Culotta, A. McCallum, and J. Betz, "Integrating probabilistic extraction models and data mining to discover relations and patterns in text," in Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006, pp. 296-303.
[6] N. Kambhatla, "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations," in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, 2004, p. 22.
[7] M. Banko, O. Etzioni, and T. Center, "The Tradeoffs Between Open and Traditional Relation Extraction," in ACL, 2008, pp. 28-36.
[8] C. C. Xavier, V. L. S. de Lima, and M. Souza, "Open information extraction based on lexical semantics," Journal of the Brazilian Computer Society, vol. 21, p. 4, 2015.
[9] M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, "Open language learning for information extraction," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 523-534.
[10] S. Soderland, B. Roof, B. Qin, S. Xu, and O. Etzioni, "Adapting open information extraction to domain-specific relations," AI Magazine, vol. 31, pp. 93-102, 2010.
[11] P. Gamallo and M. Garcia, "Multilingual open information extraction," in Portuguese Conference on Artificial Intelligence, 2015, pp. 711-722.
[12] V. Reshadat, M. Hoorali, and H. Faili, "A Hybrid Method for Open Information Extraction Based on Shallow and Deep Linguistic Analysis," Interdisciplinary Information Sciences, vol. 22, pp. 87-100, 2016.
[13] L. Del Corro and R. Gemulla, "ClausIE: clause-based open information extraction," in Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 355-366.
[14] A. Culotta and A. McCallum, "Confidence estimation for information extraction," in Proceedings of HLT-NAACL 2004: Short Papers, 2004, pp. 109-112.
[15] A. McCallum and D. Jensen, "A note on the unification of information extraction and data mining using conditional-probability, relational models," 2003.
[16] F. Wu and D. S. Weld, "Open information extraction using Wikipedia," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 118-127.
[17] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam, "Open Information Extraction: The Second Generation," in IJCAI, 2011, pp. 3-10.
[18] L. Qiu and Y. Zhang, "Zore: A syntax-based system for chinese open relation extraction," in Proceedings of EMNLP, 2014.
[19] Y.-H. Tseng, L.-H. Lee, S.-Y. Lin, B.-S. Liao, M.-J. Liu, H.-H. Chen, et al., "Chinese open relation extraction for knowledge acquisition," EACL 2014, p. 12, 2014.
[20] C. Castella Xavier, S. de Lima, V. Lúcia, and M. Souza, "Open information extraction based on lexical-syntactic patterns," in Intelligent Systems (BRACIS), 2013 Brazilian Conference on, 2013, pp. 189-194.
[21] P. Cimiano and J. Wenderoth, "Automatically learning qualia structures from the web," in Proceedings of the ACL-SIGLEX workshop on deep lexical acquisition, 2005, pp. 28-37.
[22] A. Akbik and J. Broß, "Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns," in WWW Workshop, 2009.
[23] A. Akbik and A. Löser, "Kraken: N-ary facts in open information extraction," in Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, 2012, pp. 52-56.
[24] F. Mesquita, J. Schmidek, and D. Barbosa, "Effectiveness and efficiency of open relation extraction," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, vol. 500, pp. 447–457, 2013.
[25] H. Bast and E. Haussmann, "Open information extraction via contextual sentence decomposition," in Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on, 2013, pp. 154-159.
[26] H. Bast and E. Haussmann, "More informative open information extraction via simple inference," in Advances in information retrieval, ed: Springer, 2014, pp. 585-590.
[27] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1535-1545.
[28] J. Christensen, S. Soderland, and O. Etzioni, "An analysis of open information extraction based on semantic role labeling," in Proceedings of the sixth international conference on Knowledge capture, 2011, pp. 113-120.
[29] V. Punyakanok, D. Roth, and W.-t. Yih, "The importance of syntactic parsing and inference in semantic role labeling," Computational Linguistics, vol. 34, pp. 257-287, 2008.
[30] R. Johansson and P. Nugues, "The effect of syntactic representation on semantic role labeling," in Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 2008, pp. 393-400.
[31] M. H. Kim and P. Compton, "Improving open information extraction for informal web documents with ripple-down rules," in Knowledge Management and Acquisition for Intelligent Systems, ed: Springer, 2012, pp. 160-174.
[32] P. Gamallo, M. Garcia, and S. Fernández-Lanza, "Dependency-based open information extraction," in Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, 2012, pp. 10-18.
[33] P. G. Otero and I. G. López, "A grammatical formalism based on patterns of part of speech tags," International journal of corpus linguistics, vol. 16, pp. 45-71, 2011.
[34] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction for the web," in IJCAI, 2007, pp. 2670-2676.
[35] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in Advances in Intelligent Data Analysis, ed: Springer, 2001, pp. 309-318.
[36] D. Downey, O. Etzioni, and S. Soderland, "A probabilistic model of redundancy in information extraction," DTIC Document2006.
[37] E. Agichtein, "Confidence estimation methods for partially supervised relation extraction," in Proc. of SIAM Intl. Conf. on Data Mining (SDM06), 2006.
[38] F. Mesquita, "Clustering techniques for open relation extraction," in Proceedings of the on SIGMOD/PODS 2012 PhD Symposium, 2012, pp. 27-32.
[39] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in Proceedings of the eighth ACM international conference on Web search and data mining, 2015, pp. 399-408.

[40] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2011.

Vahideh Reshadat is a Ph.D. student of ICT Department of Malek-Ashtar University of Technology in Tehran, Iran. She obtained her B.Sc. degree in software engineering in 2008, and her M.Sc. degree in software engineering in 2011. Her main research area includes Natural Language Processing field and specially Information Extraction, Query Expansion and Ontology Learning.

Maryam Hourali was born in Shahrood, Iran. She received the Ph.D. degree from Tarbiat Modares University of Technology in 2012. Her main research area includes natural language processing field and specially text summarization, ontology learning and text analysis.

Heshaam Faili received his B.Sc. degree and M.Sc. degree in Software Engineering from Sharif University of Technology in 1997 and 1999 and his Ph.D. degree in Artificial Intelligence from the same university in 2006. At present, he is an Associate Professor of University of Tehran. His areas of research include Machine Intelligence and Robotics, Information Technology, Software.