Appl. Math. Inf. Sci. 6 No. 2S pp. 627S-633S (2012)

Applied Mathematics & Information Sciences
An International Journal

Joint Inference: a Statistical Approach for Open Information Extraction

Yongbin Liu and **Bingru Yang**
School of Computer and Communication Engineering, University of Science & Technology of Beijing, 100083 Beijing, China
Corresponding author: Email: qingbinliu@163.com

Received June 23, 2010; Revised March 25, 2011; Accepted 18 June 2011
Published online: 1 January 2012

Abstract: In recent decades, natural language processing has great progress. Better model of each sub-problem achieves 90% accuracy or better, such as part-of-speech tagging and phrase chunking. However, success in integrated, end-to-end natural language understanding remains elusive. It is mainly due to the systems processing sensory input typically have a pipeline architecture: the output of each stage is the input of the next, errors cascade and accumulate through the pipeline of naively chained components, and there is no feedback from later stages to earlier ones. Actually, later stages can help earlier ones to process. For example, the part-of-speech tagger needs more syntactic and semantic information to make this choice. Previous researches tend to ignore this. So errors in each step will propagate to later ones. In current, a number of researchers have paid attention to this problem and proposed some joint approaches. But they do not perform Open Information Extraction (Open IE), which can identify various types of relations without requiring pre-specifications. In this paper, we propose a statistical modeling such unified consideration known as joint inference, which is based on Markov logic and can perform both traditional relation extraction and Open IE. The proposed modeling significantly outperforms the other Open IE systems in terms of both precision and recall. The joint inference is efficient and we have demonstrated its efficacy in real-world Open IE detection tasks.

Keywords: Natural Language Processing, Open Information Extraction, Markov logic, Joint Inference.

1 Introduction

Compare with past years, a great progress has been made in some important subtasks of natural language processing, such as POS (part of speech) Tagging, Word Segmentation, Syntactic Analysis, Named Entity Recognition, Entity-relationship Recognition, Semantic Analysis and Coreference Resolution. And the accuracy of some better model for subtask also was 90% or better [1]. But in recent years, some researchers hold that isolated processing these subtasks is improper. For example, when we process Entity Recognition subtask and Entity Analysis subtask, the Entity is interdependence rather than isolated, so it is difficulty to perform the two subtasks independently. In some cases entity recognition can address the problem of entity analysis directly, and in some cases, entity analysis can improve the accuracy of entity recognition. Information extraction is a fast developing branch of natural language processing, thus attentions should be paid to the association of extracted data. The data of information extraction often are prepared for knowledge discovery. However, relationships and rules between the extracted data were usually neglected, so the following knowledge discovery doesn't dig out some potential rules and knowledge hidden under the individual data.

To resolve these issues, there are some joint relation models for information extraction proposed, such as an integrated Model [2] and statistics relation model [3]. But these relation models are all aiming at information extraction

about special domain, which aren't suitable for Open Information Extraction (OpenIE) [4]. Therefore, we introduce Markov Logic Networks model [5] to information extraction domain, and propose a new model to accomplish joint inference for OpenIE. The method is a statistical approach based Markov Logic Networks and realizes a joint inference on the several stages of OpenIE.

The reminder of this paper is organized as follows: section 2 introduces related work. Section 3 describes Markov Logic details for the method including inference and learning. Section 4 introduces our model in terms of MLNs (section 3) and illustrate how to inference. Section 5 introduces the experiments on two datasets. Section 6 gives the conclusions and discussions.

2 Related Work

Many researchers have explored issues of joint inference in natural language processing. McCallum and Jensen had advocated the use of joint probabilistic models in natural language processing [1]. In order to avoid cascading errors, Finkel et al. presented approximate bayesian inference for linguistic annotation pipelines [6]. The feed-forward approach to inference is a classic method in Bayesian networks, but has the drawback that it only allows information to flow in one direction and can not achieve bi-direction. Then, Wellner et al. described an approach to integrated inference for extraction and coreference based on conditionally-trained undirected graphical models [2]. This approach allows information to flow in bi-directions, but it requires a restrictive approximation for the full distribution of large-output components.

Recently, Markov logic became a hot framework for joint model. Pedro Domingos et al. [3,7] presented joint inference in information extraction, where segmentation of all records and entity resolution are performed together in a single integrated inference process. But it focus on narrow and pre-specified requests from small homogeneous corpora (e.g., specific to citation matching), can not applicable to other information extraction tasks (e.g., extraction from Web pages, or from free text). Although, Michele Banko et al. introduce a approach to Open Information Extraction from the Web [4]. The approach still is a pipeline architecture. Some researchers may be inspired to resolve the problem by Hoifung Poon and Pedro Domingos. For example, Wanxiang Che and Ting Liu [8] proposed a Markov logic model that jointly labels semantic roles and disambiguates all word senses; Ivan Meza-Ruiz and Sebastian Riedel [9] introduced jointly model in identifying predicates, arguments and senses using Markov Logic. But, they all can not implement bi-directional joint inference. Based on the above issues, we propose a joint inference model based on Markov Logic Networks. It not only applies to the Open Information Extraction, but also achieves bi-directional joint inference.

3 Markov Logic

3.1 Markov Logic Networks

Markov Logic [5] is a Statistical Relational Learning language based on First Order Logic and Markov Networks. Markov logic can be understood as a knowledge representation with a weight attached to a first-order logic formula. A first-order KB (Knowledge Base) can be seen as a set of hard constraints on the set of possible worlds: if a world violates even one formula, it has zero probability[5]. The basic idea in Markov logic is to soften these constraints: when a world violates one formula in the KB it is less probable, but not impossible[5]. The fewer formulas a world violates, the more probable it is. Each formula has an associated weight that reflects how strong a constraint it is: the higher the weight, the greater the difference in log probability between a world that satisfies the formula and one that does not, other things being equal [5].

In Markov Logic a set of weighted formulae is called a Markov Logic Networks (MLNs). In a set of pairs (F_i, w_i), F_i is a first-order formula and w_i is the real weight of the formula. It assigns the probability [5]:

$$P(X=x)=\frac{1}{Z}\exp\left(\sum_i w_i n_i(x)\right)=\frac{1}{Z}\prod_i \phi_i(x_{\{i\}})^{n_i(x)} \qquad (3.1)$$

Where $n_i(x)$ is the number of true groundings of F_i in x, $x_{\{i\}}$ is the state (truth values) of the atoms appearing in F_i, and $\phi_i(x_{\{i\}})=e^{w_i}$.

3.2 Inference and Learning

Maximum a posteriori (MAP) inference in Markov networks involves finding the most likely state of a set of hidden ground atoms y given the state of a set of observed ground atoms x. The

inference tasks are defined as formula (3.2)[5]:

$$\arg\max_{y} P(y|x) = \arg\max_{y} \frac{1}{Z_x} \exp(\sum_i w_i n_i(x,y)) = \arg\max_{y} \sum_i w_i n_i(x,y) \quad (3.2)$$

Z_x is normalization constant. w_i is the weight of the ith formula, and n_i is the number of satisfied groundings. Therefore, maximal a posteriori probability (MAP) problem is summarized as to find a value which can make the sum of the weights of clauses is maximum value. The most widely used approximate solution to this problem is Markov chain Monte Carlo (MCMC) [5]. However, this is too slow for arbitrary formulas. We apply a method that is both exact and efficient: Cutting Plane Inference(CPI) [10] with Integer Linear Programming (ILP) as base solver. CPI incrementally solves partial Ground Markov Networks, adding formulae only if they are violated in the current solution[10]. CPI begins with a subset of factors/edges and solves the MAP problem for this subset using the base solver [9].

We learn the weights associated with each MLNs using Online Max-Margin Weight Learning method [11].

4 Model

4.1 Predicate Definition

Extracting relational tuples from Open Information is the goal we are pursuing. Extractions take the form of a tuple $T=(e_i, r_{i,j}, e_j)$ [4], $i<j$, where e_i and e_j are strings meant to denote entities , and $r_{i,j}$ is a string meant to denote a relationship between them. In the tuple, e_i and e_j are base noun phrases that do not contain nested noun phrases, or optional phrase modifiers, such as prepositional phrases [4]. For example, the forms of the tuple follow as[4]:

(<proper noun>, acquired, <proper noun>)
(<proper noun>, was born in, <proper noun>)
(<proper noun>, go to, <proper noun>)
(<proper noun>, come from, <proper noun>)
(<proper noun>, become, <noun phrase>)
(<proper noun>, studied at, <proper noun>)
(<proper noun>, convert, <proper noun>)
(<proper noun>, derive from, < proper noun >)
(<proper noun>, was founded by, <proper noun>)
(<proper noun>, worked in, <proper noun>)
⋮

In traditional Open IE system, e_i and e_j that denoted entities were first extracted, and then $r_{i,j}$ that denoted a relationship between them was extracted. In fact, $r_{i,j}$ denotes the implication of predicate, e_i and e_j are the semantic roles of the predicate, which make it easier for identifying the predicate. So the first task of our model is the extraction $predicate(r_{i,j})$.

Then the joint inference based on MLNs for Open Information Extraction is described in detail. Conceptually we divide our system into two stages: one stage arms to identify the predicates of a sentence, the other stage is responsible for identifying the associated e_i and e_j of these predicates.

In our joint inference model, seven hidden predicates are defined for Open IE. For predicate identification, we use two predicates $isPredicate(p)$ and $isRelation(p,t)$. $isPredicate(p)$ indicates that the word in the position p is a predicate. $isRelation(p,t)$ indicates that the word in the position p and the word in the position t which is preposition constitute a relationship r. For example, $T=(<proper\ noun>, graduated\ from, <proper\ noun>)$, string "graduated from" is the relationship r. For e_i and e_j identification, the predicates are used as follows: $isEntity(i,j)$, $hasRelation(p,e_i)$, $preRelation(p,e_i)$, $sucRelation(p,e_i)$ and $isTuple(e_i,r_{i,j},e_j)$. The predicate $isEntity(i,j)$ signals that the words from the position i to j are an entity of some (unspecified) predicate. The predicate $hasRelation(p,e_i)$ indicates that the entity e_i at position i is a semantic role of the predicate in position p. The predicate $preRelation(p,e_i)$ corresponds to the decision that e_i at position i has the role predecessor with respect to the predicate in position p. The predicate $sucRelation(p,e_i)$ corresponds to the decision that e_i at position i has the role successor with respect to the predicate in position p. The predicate $isTuple(e_i,r_{i,j},e_j)$ corresponds to a tuple $T=(e_i,r_{i,j},e_j)$, $(i<j)$.

In addition to the hidden predicates, we define observable predicates to represent the information available in Open Information. The predicate $word(i,w)$ indicates that token i has word w; the predicate $pos(i,t)$ indicates that token i has POS

tag t; the predicate $lemma(i,l)$ indicates that token i has lemma l. For accomplishing a "higher bandwidth" of communication in joint inference, a predicate $Similar(s,i,j,s',i',j')$ is defined especially, which is true if sentence s and s' contain similar strings at positions i to j and i' to j', respectively.

4.2 Local formulae

If a formula groundings relate any number of observed ground atoms to exactly one hidden ground atom, the formula is local [9]. For example,

$$lemma(p,+l_1) \wedge lemma(e,+l_2) \Rightarrow has\,\mathrm{Re}lation(p,e) \quad (4.1)$$

The formula (4.1) means that if the predicate lemma at position p is l_1 and the lemma at position e is l_2, then the predicate p and the e have the semantic role with some possibility. The + notation indicates that the MLN contains one instance of the rule, with a separate weight, for each assignment of the variables with a plus sign [3].

The local formulae for $is\mathrm{P}redicate(p)$, $is\mathrm{Re}lation(p,e_i)$ and $isEntity(i,j)$ aim to capture the relation of the tokens with their lexical and syntactic surroundings. This includes formulae such as

$$pos(p,verb) \Rightarrow is\mathrm{P}redicate(p) \quad (4.2)$$

$$lemma(p,+l_1) \wedge pos(t,prep) \Rightarrow is\mathrm{Re}lation(p,t) \quad (4.3)$$

The formula (4.2) implies that if a certain token has POS tag verb the token is a predicate with a weight. The formula (4.3) means that if the predicate lemma at position p is l_1 and POS of the word at position t is preposition, $(p<t)$, then the predicate p and the preposition t constitute a relationship string with some possibility.

4.3 Global formulae

Global formulae relate several hidden ground atoms[9]. We use this type of formula for the purposes that ensure consistency between the predicates of all stages and structural constraints [9]. For example,

$$has\,\mathrm{Re}lation(p,e_i) \wedge i<p \Rightarrow pre\,\mathrm{Re}lation(p,e_i) \quad (4.4)$$

$$has\,\mathrm{Re}lation(p,e_j) \wedge j>p \Rightarrow suc\,\mathrm{Re}lation(p,e_j) \quad (4.5)$$

$$pre\,\mathrm{Re}lation(p,e_i) \wedge suc\,\mathrm{Re}lation(p,e_j) \Rightarrow isTuple(e_i,r_{i,j},e_j) \quad (4.6)$$

Formula (4.4), formula (4.5) and formula (4.6) imply that if the e_i and e_j have the roles with the predicate in position p $(i<p<j)$, $isTuple(e_i,r_{i,j},e_j)$ can be derived with a weight. For full use of output of every stage, we add the following formula set:

$$isTuple(e_i,r_{i,j},e_j) \Rightarrow pre\,\mathrm{Re}lation(p,e_i) \quad (4.7)$$

$$isTuple(e_i,r_{i,j},e_j) \Rightarrow suc\,\mathrm{Re}lation(p,e_j) \quad (4.8)$$

$$Similar(s,i,j,s',i',j') \wedge isEntity(i,j) \Rightarrow isEntity(i',j') \quad (4.9)$$

$$Similar(s,p,s',p') \wedge is\,\mathrm{P}redicate(p) \Rightarrow is\,\mathrm{P}redicate(p') \quad (4.10)$$

The formulae (4.7)-(4.10) can achieve information to flow in bi-direction. Meanwhile, we will refer to formulae that serve the first purpose as structural constraints. For example, a structural constraint is given by the (Unique) formula (4.11).

$$pre\,\mathrm{Re}lation(p,e_1) \wedge e_1 \neq e_2 \Rightarrow \neg pre\,\mathrm{Re}lation(p,e_2) \quad (4.11)$$

The same unique constraint also happens on the $suc\mathrm{Re}lation(p,e_i)$.

$$suc\,\mathrm{Re}lation(p,e_1) \wedge e_1 \neq e_2 \Rightarrow \neg suc\,\mathrm{Re}lation(p,e_2) \quad (4.12)$$

The formulae (4.11) and (4.12) express that each entity appears only once for a predicate and each the semantic role of a predicate should be labeled with one and only one label.

As we will see in the experimental section, this joint inference suffices to outperform the state of the art in Open IE.

5 Experiments

5.1 Experiments Setting

Our experiments are performed on two datasets. The first one is the OntoNotes Release 3.0 corpus [12]. It aims to annotate a large corpus comprising various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows) in three languages (English, Chinese, and Arabic) with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference)[12]. We will do our experiments on the OntoNotes 3.0 English datasets which have been split into three

sections: weblogs , broadcast news , and magazine. The second corpus is built from the web crawler. To remove noise, such as page heads, navigation bars, etc., we first partition the crawled webpages into blocks using a visual parser [13]. The blocks in the center of a webpage are selected to compose our dataset. All the text sentences in the blocks are parsed using a part-of-speech tagger to get the POS tagging results [14]. We collect 3 million such blocks and will refer to this dataset as W3M. For two datasets, we use the Conditional Random Field (CRF) model annotated part-of-speech. The lemma of each word is extracted using WordNet tool.

To evaluate our joint inference model, we compare precision, recall and F_1 of our joint inference with that of an Open IE system that is pipeline model on a set of relations. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. F_1 is the harmonic mean of the recall and precision. Recall expresses the ability of the system to retrieve all instances; precision expresses the correct rate of system in retrieved instances. Precision and recall from two different sides comprehend the system performance. F_1 is a index of the combined precision and recall, because in the different systems for some tasks the value of precision and recall is not easy to directly compare. The use of F_1 can be more intuitive the expression of system performance.

5.2 Results

The performances of these systems on OntoNotes dataset are shown in Table 5.1. Table 5.2 shows the results of these systems on the W3M dataset.

Table 5.1: Here we show the results of the comparison for the OntoNotes dataset. Pipeline refers to the Pipeline system, Joint to our Joint Inference system. P refers to the precision, R to the recall.

Categories		*isPredicate*	*isEntity*	*hasRelation*	*isTuple*
Pipeline	P	0.93	0.91	0.87	0.86
	R	0.88	0.85	0.79	0.70
	F_1	0.90	0.87	0.82	0.77
Joint	P	0.93	0.91	0.89	0.89
	R	0.90	0.87	0.83	0.81
	F_1	**0.91**	**0.88**	**0.85**	**0.84**

From Table 5.1, it can be seen that the joint inference performs better than the Pipeline model on the predicate identification, entity identification, the identification of semantic relation of predicate and entity, extracting relational tuples. Note that entity identification are more difficult than the predicate identification; because the composition of entity is more complex.

Table 5.2: shows the results of the comparison for the W3M dataset. Pipeline refers to the Pipeline system, Joint to our Joint Inference system. P refers to the precision, R to the recall.

Categories		*isPredicate*	*isEntity*	*hasRelation*	*isTuple*
Pipeline	P	0.83	0.80	0.75	0.73
	R	0.80	0.77	0.70	0.65
	F_1	0.81	0. 78	0.72	0.68
Joint	P	0.89	0.85	0.82	0.80
	R	0.88	0.83	0.76	0.73
	F_1	**0.88**	**0.83**	**0.78**	**0.76**

From Table 5.2, it obviously shows that the joint inference performs much better than the Pipeline model on the predicate identification, entity identification, the identification of semantic relation of predicate and entity, extracting relational tuples. Meanwhile, the result on the OntoNates dataset is better than on the W3M dataset. This is due to the W3M dataset is more open types of relations without requiring pre-specifications. Whether on OntoNotes dataset or on the W3M dataset, the joint inference model is better effective than the pipeline model for Open IE.

6 Conclusions and Discussions

This paper presents a joint inference system for Open IE. It is a statistical approach that uses the general relational model—Markov logic networks (MLNs), which can be configured to perform different levels of relation extraction and can perform joint inference. Finally, we demonstrate the benefit of our joint inference on open large scale data when compared to a pipeline system.

The joint inference opens broad ways for future improvements and extensions. Currently, we apply a CRF method to get the POS tagging results. In the future, we plan to integrate our

model with the POS tagging method.

Acknowledgments

We would like to thank National Natural Science Foundation of China and Science and Technology Project of Bejing, China. This work is supported in Projects 60875029 and 61175048 by NSFC.

References

[1] Andrew McCallum, Joint Inference for Natural Language Processing. Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), (2009), 1.

[2] Ben Wellner, Andrew Mccallum, Fuchun Peng, Michael Hay. An integrated, conditional model of information extraction and coreference with application to citation matching. Proceeding UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence, (2004), 593 -601.

[3] Hoifung Poon and Pedro Domingos, Joint Inference in Information Extraction. AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence, (2007), 913-918

[4] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, Open information extraction from the web. Twentieth International Joint Conference on Artificial Intelligence, (2007), 2670-2676.

[5] M. Richardson and P. Domingos, Markov logic networks, Machine Learing. Vol.62, No.1-2, (2006), 107-136.

[6] Finkel, J.R., Manning, C.D., Ng, A.Y., Solving the problem of cascading errors: Approximate bayesian inference for linguistic annotation pipelines. Conference on Empirical Methods on Natural Language Processing (EMNLP), (2006).

[7] Parag Singla and Pedro Domingos, Entity Resolution with Markov Logic. Proceedings of the Sixth International Conference on Data Mining (ICDM'06), (2006).

[8] Wanxiang Che and Ting Liu, Jointly Modeling WSD and SRL with Markov Logic. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), (2010), 161-169.

[9] Sebastian Riedel and Ivan Meza-Ruiz, Collective Semantic Role Labelling with Markov Logic. Proceedings of the 12th Conference on Computational Natural Language Learning, (2008), 193-197.

[10] Sebastian Riedel, Improving the accuracy and efficiency of map inference for markov logic. In UAI '08: Proceedings of the Annual Conference on Uncertainty in AI, (2008).

[11] Tuyen N. Huynh Raymond J. Mooney, Online Max-Margin Weight Learning for Markov Logic Networks. Proceedings of the Eleventh SIAM International Conference on Data Mining (SDM11), (2011), 642-651.

[12] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp ?catalogId=LDC2009T24.

[13] Jun Zhu, Zaiqing Nie, Xiaojing Liu, Bo Zhang and Ji-Rong Wen, StatSnowball: a Statistical Approach to Extracting Entity Relationships. 18th International World Wide Web Conference, (2009), 101-110.

[14] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma, Simultaneous record detection and attribute labeling in web data extraction. Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2006).

[15] Yang Chuan and Li Chen, The Application of Support Vector Machine in the Modeling of the Hysteresis in Sensor. Advanced Science Letters, Vol.4, No.4-5, (2011), 1371-1375.

[16] Bingru Yang and WeiHou, KAAPRO:An approach of protein secondary structure prediction based on KDD* in the compound pyramid prediction model, Expert Systems With Applications, Vol.36, No.1, (2009), 9000-9006.

[17] Chunping Ouyang, Changjun Hu and Zhenyu Liu, Data Grid and GIS Technology for E- Science Application: A Case Study of Gas Network Safety Evaluation. Proceedings of the Fourth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2010), IEEE Computer Society, (2010), 520-525.

[18] A.H.Abd Ellah, Parametric prediction limits for generalized exponential distribution using record observations. Applied Mathematics & Information Sciences, Vol.3, No.2, (2009), 135–149.

[19] Hoifung Poon and Pedro Domingos, Unsupervised Semantic Parsing. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09), (2009).

[20] Imran Sarwar Bajwa, Context Based Meaning Extraction by Means of Markov Logic. International

Journal of Computer Theory and Engineering, Vol. 2, No. 1, February, (2010) 1793-8201.

[21] Eugene Agichtein and Luis Gravano, Snowball: Extracting Relations from Large Plain-Text Collections. International Conference on Digital Libraries, (2000).

[22] Kristina Toutanova, Aria Haghighi, and Christopher D. Manning，Joint learning improves semantic role labeling. In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, (2005).

Yongbin Liu is a Ph.D student at School of Computer and Communication Engineering, University of Science & Technology of Beijing, Beijing, China. His research interests are in the areas of machine learning and natural language processing.

Prof. Bingru Yang serves in University of Science and Technology Beijing as a life chief professor, principal level, Ph.D. supervisor of School of Computer and Communication Engineering and dean of Institute of Knowledge Engineering, who once severs as vice director of Computer Department and sub decanal of School of Information Engineering. Prof. Yang has published 550 pieces of paper in internal famous journals such as International Journal on Artificial Intelligence Tools, Fuzzy Sets and Systems, Acta Physica Sinica, Expert Systems with Applications etc. Prof. Bingru Yang has always been engaging in teaching and researching in the field of modern mathematics and computer science and technology. Main contribution in scientific research and teaching is widely evaluated as two original theory systems –knowledge discovery theory based on inner cognitive mechanism (KDTICM) and KM Teaching Methodology.