Collective Semantic Role Labeling on Open News Corpus by Leveraging Redundancy

[1,2]Xiaohua Liu, [3]Kuan Li*, [4]Bo Han*, [2]Ming Zhou,
[2]Long Jiang, [5]Daniel Tse* and [3]Zhongyang Xiong
[1]School of Computer Science and Technology
Harbin Institute of Technology
[2]Microsoft Research Asia
[3]College of Computer Science
Chongqing University
[4]School of Software
Dalian University of Technology
[5]School of Information Technologies
The University of Sydney
{xiaoliu, v-kuli, v-bohan, mingzhou, longj}
@microsoft.com
dtse6695@it.usyd.edu.au
zyxiong@cqu.edu.cn

Abstract

We propose a novel MLN-based method that collectively conducts SRL on groups of news sentences. Our method is built upon a baseline SRL, which uses no parsers and leverages redundancy. We evaluate our method on a manually labeled news corpus and demonstrate that news redundancy significantly improves the performance of the baseline, e.g., it improves the F-score from 64.13% to 67.66%.

1 Introduction

Semantic Role Labeling (SRL, Màrquez, 2009) is generally understood as the task of identifying the arguments of a given predicate and assigning them semantic labels describing the roles they play. For example, given a sentence *The luxury auto maker sold 1,214 cars.*, the goal is to identify the arguments of *sold* and produce the following output: [A0 *The luxury auto maker*] [V *sold*] [A1 *1,214 cars*]. Here *A0* represents the *seller*, and *A1* represents the things *sold* (CoNLL 2008 shared task, Surdeanu et al., 2008).

Gildea and Jurafsky (2002) first tackled SRL as an independent task, which is divided into several sub-tasks such as argument identification, argument classification, global inference, etc. Some researchers (Xue and Palmer, 2004; Koomen et al., 2005; Cohn and Blunsom, 2005; Punyakanok et al., 2008; Toutanova et al., 2005; Toutanova et al., 2008) used a pipelined approach to attack the task. Some others resolved the sub-tasks simultaneously. For example, some work (Musillo and Merlo, 2006; Merlo and Musillo, 2008) integrated syntactic parsing and SRL into a single model, and another (Riedel and Meza-Ruiz, 2008; Meza-Ruiz and Riedel, 2009) jointly handled all sub-tasks using Markov Logic Networks (MLN, Richardson and Domingos, 2005).

All the above methods conduct sentence level SRL, and rely on parsers. Parsers have showed great effects on SRL performance. For example, Xue and Palmer (2004) reported that SRL performance dropped more than 10% when they used syntactic features from an automatic parser instead of the gold standard parsing trees. Even worse, parsers are not robust and cannot always analyze any input, due to the fact that some inputs are not in the language described by the parser's formal grammar, or adequately represented within the parser's training data.

* This work has been done while the author was visiting Microsoft Research Asia.

Coling 2010: Poster Volume, pages 725–729,
Beijing, August 2010

We propose a novel MLN-based method that collectively conducts SRL on groups of news sentences to leverage the content redundancy in news. To isolate the negative effect of noise from parsers and thus focus on the study of the contribution of redundancy to SRL, we use no parsers in our approach. We built a baseline SRL, which depends on no parsers, and use the MLN framework to exploit redundancy. Our intuition is that SRL on one sentence can help that on other differently phrased sentences with similar meaning. For example, consider the following sentence from a news article:

> *A suicide **bomber** blew himself up Sunday in market in Pakistan's **northwest** crowded with shoppers ahead of a Muslim holiday, **killing** 12 people, including a mayor who once supported but had turned against the Taliban, officials said.*

The state-of-art MLN-based system (Meza-Ruiz and Riedel, 2009), hereafter referred to as MLNBS for brevity, incorrectly labels *northwest* instead of *bomber* as *A0* of *killing*. Now consider another sentence from another news article:

> *Police in northwestern Pakistan say that a suicide **bomber** has **killed** at least 13 people and wounded dozens of others.*

Here MLNBS correctly identify *bomber* as A0 of *killing*. When more sentences are observed where *bomber* as *A0* of *killing* is correctly identified, we will be more confident that *bomber* should be labeled as *A0* of *killing*, and that *northwest* should not be the *A0* of *killing* according to the constraint that one predicate has at most one A0.

We manually construct a news corpus to evaluate our method. In the corpus, semantic role information is annotated and sentences with similar meanings are grouped together. Experimental results show that news redundancy can significantly improve the performance of the baseline system.

Our contributions can be summarized as follows:

1. We present a novel method that conducts SRL on a set of sentences collectively, instead of on a single sentence, by extending MLNBS to leverage redundancy.
2. We show redundancy can significantly improve the performance of the baseline system, indicating a promising research direction towards open SRL.

In the next section, we introduce news sentence extraction and clustering. In Section 3, we describe our collective inference method. In Section 4, we show our experimental results. Finally, in Section 5 we conclude our paper with a discussion of future work.

2 Extraction and Clustering of News Sentences

To construct a corpus to evaluate our method, we extract sentences from clustered news articles returned by news search engines such as Bing and Google, and divide them into groups so that sentences in a group have similar meaning.

News articles in the same cluster are supposed to report the same event. Thus we first group sentences according to the news cluster they come from. Then we split sentences in the same cluster into several groups according to the similarity of meaning. We assume that two sentences are more similar in meaning if they share more synonymous proper nouns and verbs. The synonyms of verbs, like *plod* and *trudge*, are mainly extracted from the Microsoft Encarta Dictionary[1], and the proper nouns thesaurus, containing synonyms such as *U.S.* and *the United States*, is manually compiled.

As examples, below are two sentence groups which are extracted from a news cluster describing Hurricane Ida.

Group 1:

- *Hurricane Ida, the first Atlantic hurricane to target the U.S. this year, plodded yesterday toward the Gulf Coast…*
- *Hurricane Ida trudged toward the Gulf Coast…*
- …

Group 2:

- *It could make landfall as early as Tuesday morning, although it was forecast to weaken further.*

[1] http://uk.encarta.msn.com/encnet/features/dictionary/dictionaryhome.aspx

- *Authorities said Ida could make landfall as early as Tuesday morning, although it was forecast to weaken by then.*
- ...

3 Collective Inference Based on MLN

Our method includes two core components: a baseline system that conducts SRL on every sentence; and a collective inference system that accepts as input a group of sentences with preliminary SRL information provided by the baseline.

We build the baseline by removing formulas involving syntactic parsing information from MLNBS (while keeping other rules) and retraining the system using the tool and scripts provided by Riedel and Meza-Ruiz (2008) on the manually annotated news corpus described in Section 4.

A collective inference system is constructed to leverage redundancy in the SRL information from the baseline.

We first redefine the predicate *role* and treat it as observed:

predicate ***role****: Int x Int x Int x Role;*

role has four parameters: the first one stands for the number of sentence in the input, which is necessary to distinguish the sentences in a group; the other three are taken from the arguments of the *role* predicate defined by Riedel and Meza-Ruiz (2008), which denote the positions of the predicate and the argument in the sentence and the role of the argument, respectively. If the predication holds, it returns 1, otherwise 0.

A hidden predicate *final_role* is defined to present the final output, which has the same parameters as the predicate *role*:

predicate ***final_role****: Int x Int x Int x Role;*

We introduce the following formula, which directly passes the semantic role from the baseline to the final output:

role(s, p, a, +r)=> final_role (s, p, a, +r) (1)

Here *s* is the sentence number in a group; *p* and *a* denote the positions of the predicate and argument in *s,* respectively*; r* stands for the role of the argument; the "+" before the variable *r* indicates that different *r* has different weight.

Then we define another formula for collective inference:

s1≠s2^lemma(s1,p1,p_lemma)^lemma(s2,p2, p_lemma)^lemma(s1,a1,a_lemma)^lemma(s2, a2,a_lemma)^role(s2,p2,a2,+r)=>final_role (s1,p1,a1,+r) (2)

Here *p_lemma(a_lemma)* stands for the lemma of the predicate(argument), which is obtained from the lemma dictionary. This dictionary is extracted from the dataset of CoNLL 2008 shared task and is normalized using synonym dictionary described in Section 2; *lemma* is an observed predicate that states whether or not the word has the lemma.

Formula 2 encodes our basic ideas about collective SRL: given several sentences expressing similar meaning, if one sentence has a predicate *p* with an argument *a* of role *r*, the other sentences would be likely to have a predicate *p'* with an argument *a'* of role *r*, where *p'* and *a'* are the same or synonymous with *p* and *a,* respectively, as illustrated by the example in Section 1.

Besides, we also apply structural constraints (Riedel and Meza-Ruiz, 2008) to *final_role*.

To learn parameters of the collective inference system, we use *thebeast* (Riedel and Meza-Ruiz, 2008), which is an open Markov Logic Engine, and train it on manually annotated news corpus described in Section 4.

4 Experiments

To train and test the collective inference system, we extract 1000 sentences from news clusters, and group them into 200 clusters using the method described in Section 2. For every sentence, POS tagging is conducted with the OpenNLP toolkit (Jason Baldridge et al., 2009), lemma of each word is obtained through the normalized lemma dictionary described in Section 3, and SRL is manually labeled. To reduce human labeling efforts, we retrain our baseline on the WSJ corpus of CoNLL 2008 shared task and run it on our news corpus, and then edit the SRL outputs by hand.

We implement the collective inference system with the *thebeast* toolkit. Precision, recall, and F-score are used as evaluation metrics. In both training and evaluation, we follow the CoNLL 2008 shared task and regard only heads of phrases as arguments.

Table 1 shows the averaged 10-fold cross validation results of our systems and the baseline, where the third and second line report the results of using and not using Formula 1 in our collective inference system, respectively.

Systems	Pre. (%)	Rec. (%)	F-score (%)
Baseline	**69.87**	59.26	64.13
CI-1	62.99	**72.96**	67.61
CI	67.01	68.33	**67.66**

Table 1. Averaged 10-fold cross validation results (Pre.: precision; Rec.: recall).

Experimental results show that the two collective inference engines (CI-1 and CI) perform significantly better than the baseline in terms of the recall and F-score, though a little worse in the precision. We observe that predicate-argument relationships in sentences with complex syntax are usually not recognized by the baseline, but some of them are correctly identified by the collective inference systems. This, we guess, explains in large part the difference in performance. For instance, consider the following sentences in a group, where *order* and *tell* are synonyms:

- Colombia said on Sunday it will appeal to the U.N. Security Council and the OAS after Hugo *Chavez*, the fiery leftist president of neighboring Venezuela, *ordered* his army to prepare for war in order to assure peace.
- President Hugo *Chavez ordered* Venezuela's military to prepare for a possible armed conflict with Colombia, saying yesterday that his country's soldiers should be ready if the U.S. tries to provoke a war between the South American neighbors.
- Venezuelan President Hugo *Chavez told* his military and civil militias to prepare for a possible war with Colombia as tensions mount over an agreement giving U.S. troops access to Colombian military bases.

The baseline cannot label (*ordered, Chavez, A0*) for the first sentence, partially owing to the syntactic complexity of the sentence, but can identify the relationship for the second and third sentence. In contrast, the collective inference systems can identify *Chavez* in the first sentence as A0 of *order* because of its occurrence in the other sentences of the same group.

As Table 1 shows, the CI system achieves the highest F-score (67.66%), and a higher precision than the CI-1 system, indicating the effectiveness of Formula 1. Consider the above three sentences. CI-1 mislabels (*ordered, Venezuela, A1*) for the first sentence because the baseline labels it for the second sentence. In contrast, CI does not label it for the first sentence because the baseline does not and (*ordered, Venezuela, A1*) rarely occurs in the outputs of the baseline for this sentence group.

We also find cases where the collective inference systems do not but should help. For example, consider the following group of sentences:

- A Brazilian *university* expelled a woman who was heckled by hundreds of fellow students when she wore a short, pink dress to class, *taking* out newspaper ads Sunday to publicly accuse her of immorality.
- The *university* also *published* newspaper ads accusing the student, Geisy Arruda, of immorality.

The baseline has identified *(published, university, A0)* for the second sentence. But neither the baseline nor our method labels *(taking, university, A0)* for the first one. This happens because *publish* is not considered as a synonym of *take*, and thus *(published, university, A0)* in the second provides no evidence for *(taking, university, A0)* in the first. We plan to develop a context based synonym detection component to address this issue in the future.

5 Conclusions and Future Work

We present a novel MLN-based method that collectively conducts SRL on groups of sentences. To help build training and test corpora, we design a method to collect news sentences and to divide them into groups so that sentences of similar meaning fall into the same cluster. Experimental results on a manually labeled news corpus show that collective inference, which leverages redundancy, can effectively improve the performance of the baseline.

In the future, we plan to evaluate our method on larger news corpora, and to extend our method to other genres of corpora, such as tweets.

References

Baldridge, Jason, Tom Morton, and Gann. 2009. *OpenNLP*, http://opennlp.sourceforge.net/

Cohn, Trevor and Philip Blunsom. 2005. Semantic role labelling with tree conditional random fields. *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages: 169-172.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Journal of Computational Linguistics*, 28(3):245–288.

Koomen, Peter, Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2005. Generalized inference with multiple semantic role labeling systems. *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages: 181-184.

Màrquez, Lluís. 2009. *Semantic Role Labeling Past, Present and Future*, Tutorial of ACL-IJCNLP 2009.

Merlo, Paola and Gabriele Musillo. 2008. Semantic parsing for high-precision semantic role labelling. *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages: 1-8.

Meza-Ruiz, Ivan and Sebastian Riedel. 2009. Jointly Identifying Predicates, Arguments and Senses using Markov Logic. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages: 155-163.

Musillo, Gabriele and Paola Merlo. 2006. Accurate Parsing of the proposition bank. *Proceedings of the Human Language Technology Conference of the NAACL*, pages: 101-104.

Punyakanok, Vasin, Dan Roth and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Journal of Computational Linguistics*, 34(2), 257-287.

Richardson, Matthew and Pedro Domingos. 2005. Markov logic networks. *Technical Report, University of Washington,* 2005.

Riedel, Sebastian and Ivan Meza-Ruiz. 2008. Collective semantic role labelling with Markov Logic. *Proceedings of the Twelfth Conference on Computational Natural Language Learning,* pages: 193-197.

Surdeanu, Mihai, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. *Proceedings of the Twelfth Conference on Computational Natural Language Learning,* pages: 159-177.

Toutanova, Kristina, Aria Haghighi and Christopher D. Manning. 2005. Joint learning improves semantic role labeling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* pages: 589-596.

Toutanova, Kristina, Aria Haghighi and Christopher D. Manning. 2008. A global joint model for semantic role labeling. *Journal of Computational Linguistics*, 34(2), 161-191.

Xue, Nianwen and Martha Palmer. 2004. Calibrating features for semantic role labeling. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* pages: 88-94.