Semantic Role Labeling for News Tweets

[1,2]Xiaohua Liu, [3]Kuan Li[*], [4]Bo Han[*], [2]Ming Zhou,
[2]Long Jiang, [3]Zhongyang Xiong and [2]Changning Huang
[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
{xiaoliu, v-kuli, v-bohan, mingzhou, longj}
@microsoft.com
zyxiong@cqu.edu.cn
v-cnh@microsoft.com

Abstract

News tweets that report what is happening have become an important real-time information source. We raise the problem of Semantic Role Labeling (SRL) for news tweets, which is meaningful for fine grained information extraction and retrieval. We present a self-supervised learning approach to train a domain specific SRL system to resolve the problem. A large volume of training data is automatically labeled, by leveraging the existing SRL system on news domain and content similarity between news and news tweets. On a human annotated test set, our system achieves state-of-the-art performance, outperforming the SRL system trained on news.

1 Introduction

Tweets are text messages up to 140 characters. Every day, more than 50 million tweets are generated by millions of Twitter users. According to the investigation by Pear Analytics (2009), about 4% tweets are related to news[1].

We divide news related tweets into two categories: those excerpted from news articles and those not. The former kind of tweets, hereafter called news excerpt, is formally written while the latter, hereafter called news tweet, varies in style and often is not grammatically correct. To understand the proportion of news tweets, we randomly selected 1000 tweets related to news, and got 865 news tweets. Following is an example of anews tweet, containing *oh, yea,* which usually appear in spoken language, and *:-(,* an emoticon.

oh yea and Chile earthquake the earth off it's axis according to NASA and shorten the day by a wee second :-((S1)

News tweets arean important information source because they keep reporting what is happening in real time. For example, the earthquake near Los Angeles that happened on Tuesday, July 29, 2008 was first reported through news tweets only seconds later than the outbreak of the quake. Official news did not emerge about this event until four minutes later. By then, "Earthquake" was trending on Twitter Search with thousands of updates[2].

However, it is a daunting task for people to find out information they are interested in from such a huge number of news tweets, thus motivating us to conduct some kind of information

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

[1] http://blog.twitter.com/2010/02/measuring-tweets.html

[2] http://blog.twitter.com/2008/07/twitter-as-news-wire.html

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 698–706,
Beijing, August 2010

extraction such as event mining, where SRL plays a crucial role (Surdeanu et al., 2003). Considering Sentence 1, suppose the agent *earthquake* and the patient *day* for the predicate *shorten* are identified. Then it is straightforward to output the event *Chile earthquake shorten the day*, which captures the essential information encoded in this tweet.

Following Màrquez (2009), we define SRL for news tweets as the task of identifying the arguments of a given verb as predicate in a news tweet and assigning them semantic labels describing the roles they play for the predicate. To make our method applicable to general information extraction tasks, rather than only to some special scenarios such as arresting event extraction, we adopt general semantic roles, i.e., Agent(*A0*), Patient(*A1*), Location(*AM-LOC*), Temporal(*AM-TMP*),etc., instead of situation-specific roles (Fillmore et al., 2004) such as Suspect, Authorities, and Offense in an arrest frame.

Our first attempt is to directly apply the state-of-art SRL system (Meza-Ruiz and Riedel, 2009) that trained on the CoNLL 08 shared task dataset(Surdeanu et al., 2008), hereafter called SRL-BS, to news tweets. Not surprisingly, we observe its F1 score drops sharply from 75.5% on news corpus to 43.3% on our human annotated news tweets, owing much to the informal written style of news tweets.

Therefore, we have to build a domain specific SRL system for news tweets. Given the diversified styles of news tweets, building such a system requires a larger number of annotated news tweets, which are not available, and are not affordable for human labeling. We propose a novel method to automatically annotate news tweets, which leverages the existing resources of SRL for news domain, and content similarity between news and news tweets. We argue that the same event is likely to be reported by both news and news tweets, which results in content similarity between the news and news tweet. Further, we argue that the news and news tweets reporting the same event tend to have similar predicate-argument structures. We tested our assumptions on the event *Chile earthquake* that happened on Match 2nd, 2010. We got 261 news and 722 news tweets published on the same day that described this event. Sentence 2 and 3 are two examples of the news excerpts and Sentence 1 is one example of news tweets for this event.

Chile Earthquake Shortened Earth Day (S2)

Chile Earthquake Shortened Day (S3)

Obviously Sentence 1, 2 and 3 all have predicate "*shortened"* with the same A0 and A1 arguments. Our manually checking showed that in average each news tweet in those 993 samples had 2.4 news excerpts that had the same predicate-argument structures.

Our news tweet annotation approach consists of four steps. First, we submit hot queries to Twitter and for each query we obtain a list of tweets. Second, for each list of tweets, we single out news excerpts using heuristic rules and remove them from the list, conduct SRL on news excerpts using SRL-BS, and cluster them in terms of the similarity in content and predicate-argument structures. Third, for each list of tweets, we try to merge every remaining tweet into one news excerpt cluster according to its content similarity to the cluster. Those that can be put into one news group are regarded as news tweet. Finally, semantic structures of news excerpts are passed to the news tweet in the same group through word alignment.

Our domain specific SRL system is then trained on automatically constructed training data using the Conditional Random Field (CRF: Lafferty et al., 2001) learning framework. Our system is evaluated on a human labeled dataset, and achieves state-of-the-art performance, outperforming the baseline SRL-BS.

Our contributions can be summarized as follows:

1) We propose to conduct SRL for news tweets for fine grained information extraction and retrieval;
2) We present a semi-supervised learning approach to train a domain specific SRL system for news tweets, which outperforms SRL-BS and achieves the state-of-the-art performance on a human labeled dataset.

The rest of this paper is organized as follows: In the next section, we review related work. In Section 3 we detail key components of our approach. In Section 4, we setup experiments and evaluate the effectiveness of our method. Final-

ly, Section 5 concludes and presents the future work.

2 Related Work

Our related work falls into two categories: SRL on news and domain adaption.

As for SRL on news, most researchers used the pipelined approach, i.e., dividing the task into several phases such as argument identification, argument classification, global inference, etc., and conquering them individually (Xue and Palmer, 2004; Koomen et al., 2005; Cohn and Blunsom, 2005; Punyakanok et al., 2008; Toutanova et al., 2005; Toutanova et al., 2008). Exceptions to the pipelined approach exist. Màrquez et al. (2005) sequentially labeled the words according to their positions relative to an argument (i.e., inside, outside or at the beginning of it). Carreras et al. (2004) and Surdeanu et al. (2007) jointly labeled all the predicates. Vickrey and Koller(2008) simplified the input sentence by hand-written and machine learnt rules before conducting SRL. Some other approaches simultaneously resolved all the sub-tasks by integrating syntactic parsing and SRL into a single model (Musillo and Merlo, 2006; Merlo and Musillo, 2008), or by using Markov Logic Networks (MLN, Richardson and Domingos, 2005) as the learning framework (Riedel and Meza-Ruiz, 2008; Meza-Ruiz and Riedel, 2009).

All the above approaches focus on sentences from news articles or other formal documents, and depend on human annotated corpus for training. To our knowledge, little study has been carried out on SRL for news tweets.

As for domain adaption, some researchers regarded the out-of-domain data as "prior knowledge" and estimated the model parameters by maximizing the posterior under this prior distribution, and successfully applied their approach to language modeling (Bacchiani and Roark, 2003) and parsing (Roark and Bacchiani, 2003). Daumé III and Marcu (2006) presented a novel framework by defining a "general domain" between the "truly in-domain" and "truly out-of-domain".

Unlike existing domain adaption approaches, our method is about adapting SRL system on news domain to the news tweets domain, two domains that differ in writing style but are linked through content similarity.

3 Our Method

Our method of SRL for news tweets is to train a domain specific SRL on automatically annotated training data as briefed in Section 1.

In this section we present details of the five crucial components of our method, i.e., news excerpt identification, news excerpt clustering, news tweets identification, semantic structure mapping, and the domain specific SRL system constructing.

3.1 News Excerpt Identification

We use one heuristic rule to decide whether or not a tweet is news excerpt: if a tweet has a link to a news article and its text content is included by the news article, it is news excerpt, otherwise not.

Given a tweet, to apply this rule, we first extract the content link and expand it, if any, into the full link with the unshorten service[3]. This step is necessary because content link in tweet is usually shortened to reduce the total amount of characters. Next, we check if the full link points to any of the pre-defined news sites, which, in our experiments, are 57 English news websites. If yes, we download the web page and check if it exactly contains the text content of the input tweet. Figure 1 illustrates this process.

Figure 1. An illustration of news excerpt identification.

To test the precision of this approach, while preparing for the training data for the experiments, we checked 100 tweets that were identified as news excerpt by this rule to find out they all are excerpted from news.

[3] http://unshort.me

3.2 News Excerpt Clustering

Given as input a list of news excerpts concerning the same query and published in the same time scope, this component uses the hierarchical agglomerative clustering algorithm (Manning et al., 2008) to divide news excerpts into groups in terms of the similarity in content and predicate-argument structures.

Before clustering, for every news excerpt, we remove the content link and other metadata such as author, retweet marks (starting with RT @), reply marks (starting with @ immediately after the author), hash tags (starting with #), etc., and keep only the text content; then it is further parsed into tokens, POS tags, chunks and syntactic tree using the OpenNLP toolkit[4]. After that, SRL is conducted with SRL-BS to get predicate-argument structures. Finally, every news excerpt is represented as frequency a vector of terms, including tokens, POS tagger, chunks, predicate-argument structures, etc. A news cluster is regarded as a "macro" news excerpt and is also represented as a term frequency vector, i.e., the sum of all the term vectors in the cluster. Noisy terms, such as numbers and predefined stop words are excluded from the frequency vector. To reduce data sparseness, words are stemmed by Porter stemmer (Martin F. Porter, 1980).

The cosine similarity is used to measure the relevance between two clusters, as defined in Formula 1.

$$CS(C,C') = \frac{CV \cdot CV'}{\|CV\| \times \|CV'\|} \quad (1)$$

Where *C, C'* denote two clusters, *CV, CV'* denote the term frequency vectors of *C* and *C'* respectively, and *CS(C, C')* stands for the cosine similarity between *C* and *C'*.

Initially, one news excerpt forms one cluster. Then the clustering process repeats merging the two most similar clusters into one till the similarity between any pair of clusters is below a threshold, which is experimentally set to 0.7 in our experiments.

During the training data preparation process, we randomly selected 100 clusters, each with 3.2 pieces of news in average. For every pair of news excerpts in the same cluster, we checked if they shared similar contents and semantic structures, and found out that 91.1% were the cases.

[4] http://opennlp.sourceforge.net/

3.3 News Tweets Identification

After news excerpts are identified and removed from the list, every remaining tweet is checked if it is a news tweet. Here we group news excerpts and news tweets together in two steps because 1) news excerpts count for only a small proportion of all the tweets in the list, making our two-step clustering algorithm more efficient; and 2) one-step clustering tends to output meaningless clusters that include no news tweets.

Intuitively, news tweet, more often than not, have news counterparts that report similar contents. Thus we use the following rule to identify news tweets: if the content similarity between the tweet and any news excerpt cluster is greater than a threshold, which is experimentally set to 0.7 in our experiments, the tweet is a news tweet, otherwise it is not. Furthermore, each news tweet is merged into the cluster with most similar content. Finally, we re-label any news tweet as news excerpt, which is then process by SRL-BS, if its content similarity to the cluster exceeds a threshold, which is experimentally set to 0.9 in our experiments.

Again, the cosine similarity is used to measure the content similarity between tweet and news excerpt cluster. Each tweet is repressed as a term frequency vector. Before extracting terms from tweet, tweet metadata is removed and a rule-based normalization process is conducted to restore abnormal strings, say "'", into their human friendly form, say '' ''. Next, stemming tools and OpenNLP are applied to get lemmas, POS tags, chunks, etc., and noisy terms are filtered.

We evaluated the performance of this approach when preparing for the training data. We randomly sampled 500 tweets that were identified as news tweets, to find that 93.8% were true news tweets.

3.4 Semantic Structure Mapping

Semantic structure mapping is formed as the task of word alignment from news excerpt to news tweet. A HMM alignment model is trained with GIZA++ (Franz and Hermann, 2000) on all (news excerpt, news tweet) pairs in the same cluster. After word alignment is done, semantic

information attached to a word in a news excerpt is passed to the corresponding word in the news tweet as illustrated in Figure 2.

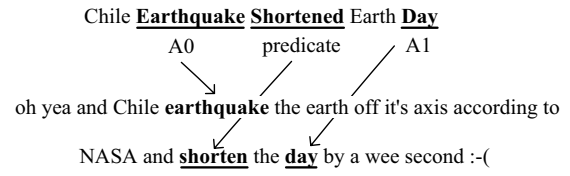

Figure 2. An example of mapping semantic structures from news excerpts to news tweets.

In Figure 2, *shorten*, *earthquake* and *day* in two sentences are aligned, respectively; and two predicate-argument structures in the first sentence, i.e., *(shortened, earthquake, A0), (shortened, day, A1)*, are passed to the second.

News tweets may receive no semantic information from related news excerpts after mapping, because of word alignment errors or no news excerpt in the cluster with similar semantic structures. Such tweets are dropped.

Mapping may also introduce cases that violate the following two structural constraints in SRL (Meza-Ruiz and Riedel, 2009): 1) one (predicate, argument) pair has only one role label in one sentence; and 2) for each predicate, each of the proper arguments *(A0~A5)* can occur at most once. Those conflicts are largely owing to the noisy outputs of SRL trained on news and to the alignment errors. While preparing for the training data for our experiments, we found 38.9% of news tweets had such conflicts.

A majority voting schema and the structural constrains are used to resolve the conflicts as described below.

1) Step 1, for every cluster, each (*predicate*, *argument*, *role*) is weighted according to its frequency in the cluster;
2) Step 2, for every cluster, detect conflicts using the structural constrains; if no conflicts exist, stop; otherwise go to Step 3;
3) Step 3, for every cluster, keep the one with higher weight in each conflicting (*predicate*, *argument*, *role*) pair; if the weights are equal, drop both;

Here is an example to show the conflicting resolution process. Consider the cluster including Sentence 1, 2 and 3, where *(shorten, earthquake, A0), (shorten, earthquake, A1), (shorten, axis, A0),* and *(shorten, day, A1)* occur 6, 4, 1 and 3 times, respectively. This cluster includes three conflicting pairs:

1) *(shorten, earthquake, A0) vs. (shorten, earthquake, A1);*
2) *(shorten, earthquake, A1) vs. (shorten, day, A1);*
3) *(shorten, earthquake, A0) vs. (shorten, axis, A0);*

The first pair is first resolved, causing *(shorten, earthquake, A0)* to be kept and *(shorten, earthquake, A1)* removed, which leads to the second pair being resolved as well; then we process the third pair resulting in *(shorten, earthquake, A0)* being kept and *(shorten, axis, A0)* dropped; finally *(shorten, earthquake, A0)* and *(shorten, day, A1)* stay in the cluster.

The conflicting resolution algorithm is sensitive to the order of conflict resolution in Step 3. Still consider the three conflicting pairs listed above. If the second pair is first processed, only *(shorten, earthquake, A0)* will be left. Our strategy is to first handle the conflict resolving which leads to most conflicts resolved.

We tested the performance of this semantic structure mapping strategy while preparing for the training data. We randomly selected 56 news tweets with conflicts and manually annotated them with SRL. After the conflict resolution method was done, we observed that 38 news tweets were resolved correctly, 9 resolved but incorrectly, and 9 remain unresolved, suggesting the high precision of this method, which fits our task. We leave it to our future work to study more advanced approach for semantic structure mapping.

3.5 SRL System for News Tweets

Following Màrquez et al. (2005), we regard SRL for tweets as a sequential labeling task, because of its joint inference ability and its openness to support other languages.

We adopt conventional features for each token defined in Màrquez et al.(2005), such as the lemma/POS tag of the current/previous/next token, the lemma of predicate and its combination with the lemma/POS tag of the current token, the voice of the predicate (active/passive), the distance between the current token and the predicate, the relative position of the current token to

the predicate, and so on. We do not use features related to syntactic parsing trees, to allow our system not to rely on any syntactic parser, whose performance depends on style and language of text, which limits the generality of our system.

Before extracting features, we perform a preprocessing step to remove tweet metadata and normalize tweet text content, as described in Section 3.3. The OpenNLP toolkit is used for feature extraction, and the CRF++ toolkit[5] is used to train the model.

4 Experiments

In this section, we evaluate our SRL system on a gold-standard dataset consisting of 1,110 human annotated news tweets and show that our system achieves the state-of-the-art performance compared with SRL-BS that is trained on news. Furthermore, we study the contribution of automatically generated training data.

4.1 Evaluation Metric

We adopt the widely used precision (Pre.), recall (Rec.) and F-score (F., the harmonic mean of precision and recall) as evaluation metrics.

4.2 Baseline System

We use SRL-BS as our baseline because of its state-of-art performance on news domain, and its readiness to use as well.

4.3 Data Preparation

We restrict to English news tweets to test our method. Our method can label news tweets of other languages, given that the related tools such as the SRL system on news domain, the word alignment tool, OpenNLP, etc., can support other languages.

We build two corpora for our experiments: one is the training dataset of 10,000 news tweets with semantic roles automatically labeled; the other is the gold-standard dataset of 1,110 news tweets with semantic roles manually labeled.

Training Dataset

We randomly sample 80 queries from 300 English queries extracted from the top stories of Bing news, Google news and Twitter trending topics from March 1, 2010 to March 4, 2010.

Submitting the 80 queries to Twitter search, we retrieve and download 512,000 tweets, from which we got 4,785 news excerpts and 11,427 news tweets, which were automatically annotated using the method described in Section 3.

Furthermore, 10,000 tweets are randomly selected from the automatically annotated news tweets, forming the training dataset, while the other 1,427 news tweets are used to construct the gold-standard dataset.

Gold-standard Dataset

We ask two people to annotate the 1,427 news tweets, following the Annotation guidelines for PropBank[6] with one exception: for phrasal arguments, only the head word is labeled as the argument, because our system and SRL-BS conduct word level SRL.

317 news tweets are dropped because of inconsistent annotation, and the remaining 1,110 news tweets form the gold-standard dataset.

Quality of Training dataset

Since the news tweets in the gold-standard dataset are randomly sampled from the automatically labeled corpus and are labeled by both human and machine, we use them to estimate the quality of training data, i.e., to which degree the automatically generated results are similar to humans'.

We find that our method achieves 75.6% F1 score, much higher than the baseline, suggesting the relatively high quality of the training data.

4.4 Result and Analysis

Table 1 reports the experimental results of our system (SRL-TS) and the baseline on the gold-standard dataset.

	Precision	Recall	F-Score
SRL-BS	36.0 %	54.5%	43.3%
SRL-TS	78.0%	57.1%	66.0%

Table 1. Performances of our system and the baseline on the gold-standard dataset.

As shown in Table 1, our system performs much better than the baseline on the gold-standard dataset in terms of all metrics. We observe two types of errors that are often made by

[5] http://crfpp.sourceforge.net/

[6] http://verbs.colorado.edu/~mpalmer/projects/ace/PB guidelines.pdf

SRL-BS but not so often by our system, which largely explains the difference in performance.

The first type of errors, which accounts for 25.3% of the total errors made by SRL-BS, is caused by the informal written style, such as ellipsis, of news tweets. For instance, for the example Sentence 1 listed in Section 1, the SRL-BS incorrectly identify earth as the A0 argument of the predicate shorten. The other type of errors, which accounts for 10.2% of the total errors made by SRL-BS, is related to the discretionary combination of news snippets. For example, consider the following news tweet:

The Chile earthquake shifted the earth's axis, "shortened the length of an Earth day by 1.26 miliseconds". (S4)

We analyze the errors made by our system and find that 12.5% errors are attributed to the complex syntactic structures, suggesting that combining our system with systems on news domain is a promising direction. For example, our system cannot identify the A0 argument of the predicate *shortened*, because of its blindness of attributive clause; in contrast, SRL-BS works on this case.

wow..the earthquake that caused the 2004 Indian Ocean tsunami shortened the day by almost 3 microseconds..what does that even mean?! HOW? (S5)

We also find that 32.3% of the errors made by our system are more or less related to the training data, which has noise and cannot fully represent the knowledge of SRL on news tweets. For instance, our system fails to label the following sentence, partially because the predicate *strike* does not occur in the training set.

8.8-Magnitude-Earthquake-Strikes-Chile (S6)

We further study how the size of automatically labeled training data affects our system's performance, as illustrated in Figure 3. We conduct two sets of experiments: in the first set, the training data is automatically labeled and the testing data is the gold-standard dataset; in the second set, half of the news tweets from the gold-standard dataset are added to the training data, the remaining half forms the testing dataset. Curve 1 and 2 represent the experimental results of set 1 and 2, respectively.

From Curve 1, we see that our system's performance increases sharply when the training data size varies from 5,000 to 6,000; then increases relatively slowly with more training data; and finally reaches the highest when all training data is used. Curve 2 reveals a similar trend.

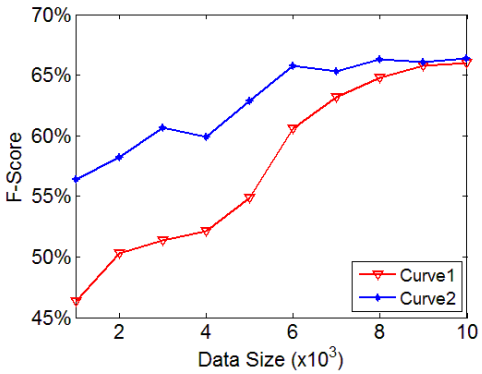

Figure 3. Performance on training data of varying size.

This phenomenon is largely due to the competing between two forces: the noise in the training data, and the knowledge of SRL encoded in the training data.

Interestingly, from Figure 3, we observe that the contribution of human labeled data is no longer significant after 6,000 automatically labeled training data is used, reaffirming the effectiveness of the training data.

5 Conclusions and Future Work

We propose to conduct SRL on news tweets for fine grained information extraction and retrieval. We present a self-supervised learning approach to train a domain specific SRL system for news tweets. Leveraging the SRL system on news domain and content similarity between news and news tweets, our approach automatically labels a large volume of training data by mapping SRL-BS generated results of news excerpts to news tweets. Experimental results show that our system outperforms the baseline and achieves the state-of-the-art performance.

In the future, we plan to enlarge training data size and test our system on a larger dataset; we also plan to further boost the performance of our system by incorporating tweets specific features such as hash tags, reply/re-tweet marks into our

CRF model, and by combining our system with SRL systems trained on news.

References

Bacchiani, Michiel and Brian Roark. 2003. Unsupervised language model adaptation. *Proceedings of the 2003 International Conference on Acoustics, Speech and Signal Processing,* volume 1, pages: 224-227

Carreras, Xavier, Lluís Màrquez, and Grzegorz Chrupała. 2004. Hierarchical recognition of propositional arguments with Perceptrons. *Proceedings of the Eighth Conference on Computational Natural Language Learning*, pages: 106-109.

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

Daumé, Hal III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1), 101-126.

Fillmore, Charles J., Josef Ruppenhofer, Collin F. Baker. 2004. FrameNet and Representing the Link between Semantic and Syntactic Relations. *Computational Linguistics and Beyond, Institute of Linguistics, Academia Sinica.*

Kelly, Ryan, ed. 2009. *Twitter Study Reveals Interesting Results About Usage*. San Antonio, Texas: Pear Analytics.

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.

Lafferty, John D., Andrew McCallum, Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pages: 282-289.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Màrquez, Lluís, Jesus Gimènez Pere Comas and Neus Català. 2005. Semantic Role Labeling as Sequential Tagging. *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages: 193-196.

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.

Och, Franz Josef, Hermann Ney. Improved Statistical Alignment Models. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages: 440-447.

Porter, Martin F. 1980. An algorithm for suffix stripping. *Program,* 14(3), 130-137.

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.

Roark, Brian and Michiel Bacchiani. 2003. Supervised and unsupervised PCFG adaptation to novel domains. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* volume 1, pages: 126-133.

Surdeanu, Mihai, Sanda Harabagiu, JohnWilliams and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics,* volume 1, pages: 8-15.

Surdeanu, Mihai, Lluís Màrquez, Xavier Carreras and Pere R. Comas. 2007. Combination strategies for semantic role labeling. *Journal of Artificial Intelligence Research,* 29(1), 105-151.

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.

Vickrey, David and Daphne Koller. 2008. Applying sentence simplification to the conll-2008 shared task. *Proceedings of the Twelfth Conference on Computational Natural Language Learning,* pages: 268-272

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.