

A Weakly-supervised Approach to Argumentative Zoning of Scientific Documents

Review by Pranjal Singh(10511)

October 6, 2013

References

- [1] Yufan Guo, Anna Korhonen, and Thierry Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 273–283, 2011.

1 Overview

The paper talks about classification of documents according to certain broad classes such as classifying a scientific literature according to objective, methodology, results obtained, etc. using weakly-supervised learning which is less expensive as compared to fully supervised approach. The paper basically investigates the performance of weakly-supervised learning for Argumentative Zoning(AZ) of scientific abstracts. AZ is an approach to information structure which provides an analysis of the rhetorical progression of the scientific argument in a document(Teufel and Moens, 2002). Because of the utility of this method in several domains(such as summarization, computational linguistics), weakly-supervised approach is much more practical.

2 Architecture

Data:Guo et al. (2010) provide a corpus of 1000 biomedical abstracts (consisting of 7985 sentences and 225785 words) annotated according to three schemes of information structure. The paper uses AZ (Mizuta et al., 2006)

Methods: Comparison between supervised classifier(SVM and CRF) with four weakly supervised classifiers: two based on semi-supervised learning (transductive SVM and semi-supervised CRF) and two on active learning (Active SVM alone and in combination with self-training).

The figure below shows the categories of AZ appearing in the corpus.

Category	Abbr.	Definition
Background	BKG	The circumstances pertaining to the current work, situation, or its causes, history, etc.
Objective	OBJ	A thing aimed at or sought, a target or goal
Method	METH	A way of doing research, esp. according to a defined and regular plan; a special form of procedure or characteristic set of procedures employed in a field of study as a mode of investigation and inquiry
Result	RES	The effect, consequence, issue or outcome of an experiment; the quantity, formula, etc. obtained by calculation
Conclusion	CON	A judgment or statement arrived at by any reasoning process; an inference, deduction, induction; a proposition deduced by reasoning from other propositions; the result of a discussion, or examination of a question, final determination, decision, resolution, final arrangement or agreement
Related work	REL	A comparison between the current work and the related work
Future work	FUT	The work that needs to be done in the future

According to Cohens kappa (Cohen, 1960) the inter-annotator agreement is relatively high: $\text{kappa}(k) = 0.85$.

3 Methodology

Feature Extraction: Documents were first identified by their features which included Zones(e.g. abstracts were divided into ten parts where zones typically occur), Words, Bi-grams, Verbs, Verb Class, POS, Grammatical Relation, Subject & Object, Voice of Verbs. These were extracted using tools such as tokenizer(detects boundaries in a sentence), C&C Tool(POS tagging, lemmatization and parsing), unsupervised spectral clustering method(to acquire verb classes). The lemma output was used to create Word, Bi-gram and Verb features.

Machine Learning Methods: SVM and CRF were used as Supervised methods on the data obtained after feature extraction. **Active SVM:** *starts with a small amount of labeled data, and iteratively chooses a proportion of unlabeled data for which SVM has less confidence to be labeled and used in the next round of learning*, **Active SVM with self-training**, **Transductive SVM:** *takes advantage of both labeled and unlabeled data* and **Semi-supervised CRF** were used as Weakly-Supervised Methods.

4 Evaluation

Results were evaluated in terms of accuracy, precision, recall and F-measure against manual AZ annotation in the corpus. 10-fold cross-validation was used to avoid bias in training data. McNemar’s test was used to measure the statistical significance between different ML methods.

5 Results

With 10% training data, ASSVM performs best with 81% accuracy and macro F-score of .76. ASVM performs with accuracy of 80% and F-score of .75. Both of them outperform supervised SVM. TSVM is the worst performing SVM-based method with an accuracy of 76% and F-score of .73 which is less than supervised SVM. But it outperforms both CRF-based methods. Only one method(ASVM) identifies six out of the seven possible categories. Other methods identify five categories. RESULTS has the highest amount of data and OBJECTIVE has the minimum. The LOCATION feature is found to be the most important feature for ASSVM. Voice, Verb class and GR contribute to general performance. Least helpful features are Word, Bi-gram and Verb because they suffer from sparse data problems. It was also found that a good performance in fully supervised experiments does not necessarily translate into a good performance in weakly-supervised experiments. The result is attached below:

Acc. F-score											
		MF	BKG	OBJ	METH	RES	CON	REL	FUT		
SVM	.77	.74	.84	.68	.71	.82	.64	-	-		
CRF	.70	.65	.75	.46	.48	.78	.76	-	-		
ASVM	.80	.75	.88	.56	.68	.87	.78	.33			
ASSVM	.81	.76	.86	.56	.76	.88	.76	-	-		
TSVM	.76	.73	.84	.61	.71	.79	.71	-	-		
SSCRF	.73	.67	.76	.48	.52	.81	.78	-	-		

6 Discussion

Almost all the methods performed as they have been performing in other domains and on work done by other authors. But TSVM didn’t perform better than SVM with the same amount of labeled data. This could be due to higher dimensional data in this work as compared to other works. SSCRf didn’t perform as expected on the data may be due to less number of labeled and unlabeled instances.

7 Future Work

The approach to active learning could be improved in various ways by experimenting with more complex query strategies such as margin sampling algorithm by (Scheffer et al., 2001) and query-by-committee algorithm by (Seung et al., 1992). Other semi-supervised methods such as Expectation Maximization algorithm could be used which has shown to outperform supervised and transductive SVM. Also, looking for other optimal features could improve the result a lot. It could also be very interesting to evaluate the usefulness of weakly-supervised identification of information structure for NLP tasks such as summarization and information extraction and for practical tasks such as manual review of scientific papers for research purposes.

Review By:

Pranjal Singh, 10511, 2013