System documentation for the IGGSA Shared Task 2016

Leonard Kriese
Universität Potsdam
Department Linguistik
Karl-Liebknecht-Straße 24-25
14476 Potsdam
leonard.kriese@hotmail.de

Abstract

This is a brief documentation of a system, which was created to compete in the shared task on *Source, Subjective Expression and Target Extraction from Political Speeches (STEPS)*. The system's model is created from supervised learning on using the provided training data and is learning a lexicon of *subjective expressions*. Then a slightly different model will be presented that generalizes a little bit from the training data.

1 Introduction

This is a documentation of a system, which has been created within the context of the IGGSA Shared Task 2016 STEPS[1]. Briefly, the main goal was to find *subjective expressions (SE)* that are functioning as opinion triggers, their *sources*, the originator of an SE, and their *targets*, the scope of an opinion. The system was aimed to perform on the domain of parliament speeches from the Swiss Parliament. The system's model was trained on the training data provided alongside the shared task and was from the same domain, preprocessed, with constituency parses from the Berkley Parser (Petrov and Klein, 2007) and had annotations of SEs and their respective targets and sources.

The model is using a mapping from grouped SEs to a set of "path-bundles", syntactic relations between SE, source and target. Since the learned SEs are a lexicon derived from the training data and are very domain-dependent, there will be a second model presented, which generalizes slightly from the training data by using the SentiWS (Remus et al., 2010) as a lexicon of SEs. There, the part-of-speech tag of each word from the SentiWS is mapped to a set of path-bundles.

[1] Source, Subjective Expression and Target Extraction from Political Speeches(STEPS)

2 System description

Participants were free in the way they could develop the system. They just had to identify subjective expressions and their corresponding target and source. Our system is using a lexical approach to find the subjective expressions and a syntactic approach in finding the corresponding target and source. First, all the words in the training data were lemmatized with the TreeTagger (Schmid, 1995), to keep the number of considered words as low as possible. Then the SE lexicon was derived from all the SEs in the training data. For each SE in the training data its path-bundle, a syntactic relation to the SE's target and source was stored. These path-bundles were derived from the constituency-parses from the Berkley Parser. For each sentence in the test data all the words were checked if they were SE candidates. If they were, their syntactic surroundings were checked as well. If these were also valid, a target and source was annotated. The test data was also lemmatized.

The outline of this paper is: the approach of deriving the syntactic relation of the SEs by introducing the concepts of "minimal trees" and "path-bundles" (Section 2.1 and Section 2.2) will be presented. Then the clustering of SEs and their path-bundles will be explained (Section 2.3) and a more generalized model (Section 2.4).

2.1 Minimal Trees

We use the term "minimal tree" for a sub-tree of a syntax tree given in the training data for each sentence with the following property: its root node is the least common ancestor of the SE, target and source[2]. From all the identified minimal trees so called path-bundles were derived. In Figure 1 and 2 you can see such minimal trees. These just focus on the part of the syntactic tree, which relates to

[2] Like the *lowest common multiple*, just for SE, target and source.

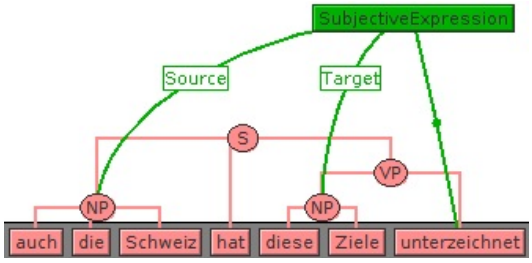

Figure 1: minimal tree covering *auch die Schweiz hat diese Ziele unterzeichnet*

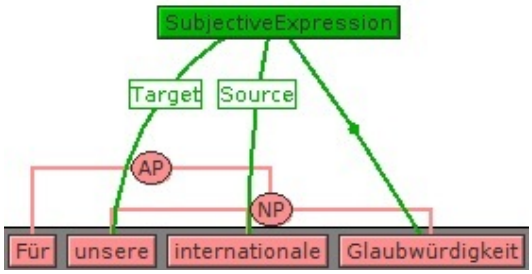

Figure 2: minimal tree covering *Für unsere internationale Glaubwürdigkeit*

SE, source and target. Following the root node of the minimal tree to the SE, target and source, path-bundles are extracted, as you can see in (1) and (2).

2.2 Path-Bundles

A path-bundle holds the paths in the minimal tree for the SE, target and source to the root node of the minimal tree.

(1) path-bundle for minimal tree in Figure 1
SE: [S, VP, VVPP]
T: [S, VP, NP]
S: [S, NP]

(2) path-bundle for minimal tree in Figure 2
SE: [NP, NN]
T: [NP, PPOSAT]
S: [NP, ADJA]

As you can see in (1) and (2), there is no distinction between terminals and non-terminals here, since SEs are always terminals (or a group of terminals) and targets and sources are sometimes one and the other. A path-bundle is expressing a syntactic relation between the SE, target and source and can be seen as a syntactic pattern. In practice many SEs have more than one path-bundle.
When the system is annotating a sentence, e.g. from the test data, and an SE is detected, the system checks the current syntax tree for one of the path-bundles an SE might have. If one bundle applies, source and target along with the SE will be annotated.
Also the flags, which appear in the training data, are stored to a path-bundle and will be annotated, when the corresponding path-bundle applies.

2.3 Clustering

After the training procedure every SE has its own set of path-bundles. To make the system more open to unseen data, the SEs were clustered in the following way: if an SE shared at least one path-bundle with another SE, they were put together into a cluster. The idea is, if SEs share a syntactic pattern towards their target and source, they are also syntactically similar and hence, should share their path-bundles. Rather than using a single SE mapped to a set of path-bundles, the system uses a mapping of a set of SEs to the set of their path-bundles.

(3) {befürworten, wünschen, nachleben, beschreiben, schützen, versprechen, verschärfen, erreichen, empfehlen, ausschreiben, verlangen, folgen, mitunterschreiben, beizustellen, eingreifen, appellieren, behandeln}

(4) {..., Regelung, Bodenschutz, Nichteintreten, Anreiz, Verteidigung, Kommentator, Kommissionsmotion, Verkehrsbelastung, Jugendstrafgesetz, Rückweisungsantrag, Konvention, Neutralitätsthematik, Europapolitik, Debatte,...}

(5) {lehnen ab, nehmen auf, ordnen an}

The clusters in (3), (4) and (5) are examples of what has been clustered in the training. This was done automatically and is presented here for illustration. As future work, we will consider manually merging some of the clusters and testing, whether that improves the performance.

2.4 Second model

The second model is generalizing a little bit from the lexicon of the training data, since the first model is very domain-dependent and should perform much worse on another domain than on the test data. The generalization is done by exchanging the lexicon learned from the training data with the words from the SentiWS (Remus et al., 2010).
This model is thus more generalized and not domain-dependent, but neither domain-specific. If

a word from the lexicon will be detected in a sentence, then all path-bundles, which begin with the same pos-tag, in the SE-path, will be considered for finding the target and source.
In general the sorting of the path-bundles is dependent from the leaf node in the SE-path, since the procedure is the following: find an SE and check if one of the path-bundles can be applied. Maybe, this can be done in a reverse way, where every node in a syntax tree is seen as a potential top-node of a path-bundle and if a path-bundle can be applied, SE, target and source will be annotated accordingly. This could be a heuristic for finding SEs without the use of a lexicon.

3 Results

In this part, the results of the two models, which ran on the STEPS 2016 data, will be presented.

Measure	Supervised	SentiWS
F1 SE exact	35.02	30.42
F1 source exact	18.29	15.62
F1 target exact	14.32	14.52
Prec SE exact	48.15	58.40
Prec source exact	27.23	34.66
Prec target exact	20.44	32.11
Rec SE exact	27.51	20.56
Rec source exact	13.77	10.08
Rec target exact	11.02	9.38

Table 1: Results of the system's runs on the main task.

Measure	Subtask A
F1 source exact	32.87
Prec source exact	36.23
Rec source exact	30.08
	Subtask B
F1 target exact	27.83
Prec target exact	37.29
Rec target exact	22.20

Table 2: Results of the system run on the subtasks.

The first system (Supervised) is the domain-dependent, supervised system with the lexicon from the training data and was the system, which was submitted to the IGGSA Shared Task 2016. The second system (SentiWS) is the system with the lexicon from the SentiWS. Speaking about Table 1, with the results for the main task, considering the F1-measure, the first system was better in finding SEs and sources but a little bit worse in finding targets.
The second system, the more general system, was better in the precision scores overall. This means, in comparison to the supervised system, that the classified SEs, targets and source were more correct. But it did no find as many as it should have found as the first system according to the recall scores. This leads to the assumption that the first system might overgenerate and is therefore hitting more of the true positives, but is also making more mistakes.
Looking at Table 2, the systemic approach is just different in terms of the lexicon of SEs and not in terms of the path-bundles. So there is no distinction between the two systems here, since all the SEs were given in the subtasks and only the learned path-bundles determined the outcome of the subtasks. For the system it seems easier to find the right sources, rather than the right targets, which is also proven by the numbers in Table 1.

4 Conclusion

In this documentation for the IGGSA Shared Task 2016 an approach was presented, which uses the provided training data. First, a lexicon of SEs was derived from the training data along with their path-bundles, indicating where to find their respective target and source. Two generalization steps were made by first, clustering SEs, which had syntactic similarities and second by exchanging the lexicon derived from the training data with a domain-independent lexicon, the SentiWS.
The first, very domain-dependent, system performed better than the more general second system according to the f-score. But the second system did not make as many mistakes in detecting SEs from the test data by looking at the precision score, so it might be worth to investigate into the direction of using a more general approach further.
The approach of deriving path-bundles from syntax trees itself is domain-independent, since it can be easily applied to any kind of parse. It would be nice to see, how the results will change, when other parsers, like a dependency parser, will be used. This is something for the future work.

References

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Tech-*

nologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

R. Remus, U. Quasthoff, and G. Heyer. 2010. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, pages 1168–1171.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer.