A Tagging Approach to Identify Complex Constituents for Text Simplification

Iustin Dornescu **Richard Evans** **Constantin Orăsan**
Research Institute in Information and Language Processing
University of Wolverhampton
United Kingdom
{I.Dornescu2, R.J.Evans, C.Orasan}@wlv.ac.uk

Abstract

The occurrence of syntactic phenomena such as coordination and subordination is characteristic of long, complex sentences. Text simplification systems need to detect and categorise constituents in order to generate simpler sentences. These constituents are typically bounded or linked by *signs of syntactic complexity*, which include conjunctions, complementisers, wh-words, and punctuation marks. This paper proposes a supervised tagging approach to classify these signs in accordance with their linking and bounding functions. The performance of the approach is evaluated both intrinsically, using an annotated corpus covering three different genres, and extrinsically, by evaluating the impact of classification errors on an automatic text simplification system. The results are encouraging.

1 Introduction

This paper presents an automatic method to determine the specific coordinating and bounding functions of several reliable signs of syntactic complexity in natural language. This method can be useful for automatic text simplification. The syntactic complexity of input text can be reduced by the application of rules triggered by patterns expressed in terms of the parts of speech of words and the syntactic linking and bounding functions of signs of syntactic complexity occurring within it (Evans, 2011). Previous work indicates that syntactic simplification can improve text accessibility (Just et al., 1996) and the reliability of NLP applications such as information extraction (Agarwal and Boggess, 1992; Rindflesch et al., 2000), machine translation (Gerber and Hovy, 1998), and syntactic parsing (Tomita, 1985; McDonald and Nivre, 2011). The research described in the current paper is part of the FIRST project[1] which aims to automatically convert documents into a more accessible form for people with autistic spectrum disorders (ASD). Many of the decisions taken in the research presented in this paper were informed by the psycholinguistic experiments carried out in the FIRST project and summarised in Martos et al. (2013).

The remainder of this paper is structured as follows. Section 2 provides background information about the context of this work, Section 3 presents the annotation scheme, Section 4 describes the approach and the main objectives of this study. The results and the main findings are presented in Section 5. Section 6 provides an overview of previous related work. In Section 7, conclusions are drawn.

2 Syntactic Simplification in the FIRST Project

Research carried out in the FIRST project and investigation of related work revealed that certain types of syntactic complexity adversely affect the reading comprehension of people with ASD (Martos et al., 2013). This section presents a brief overview of the context in which this research is carried out. It builds on the approach proposed by Evans (2011) who presented a rule-based method to simplify sentences containing coordinated constituents to facilitate information extraction. In that work, punctuation marks and conjunctions were considered to be reliable signs of syntactic complexity in English. These signs were automatically classified in accordance with a scheme indicating their specific syntactic linking function. They then serve as triggers for the application of distinct sets of simplification rules. Their accurate labelling is thus a prerequisite for

[1] http://www.first-asd.eu

Proceedings of Recent Advances in Natural Language Processing, pages 221–229,
Hissar, Bulgaria, 7-13 September 2013.

the simplification process.

In that work, signs of syntactic complexity were considered to belong to one of two broad classes, denoted as *coordinators* and *subordinators*. These groups were subcategorised according to class labels specifying the syntactic projection level of conjoins[2] and of subordinated constituents, and the grammatical category of those phrases. Manual annotation of a limited set of signs was exploited to develop a memory-based learning classifier that was used in combination with a part-of-speech tagger and a set of rules to rewrite complex sentences as sequences of simpler sentences. Extrinsic evaluation showed that the simplification process evoked improvements in information extraction from clinical documents.

One weakness of the approach presented by Evans (2011) is that the set of functions of signs of syntactic complexity was derived by empirical analysis of rather homogeneous documents from a specialised source (a collection of clinical assessment items). The restricted range of linguistic phenomena encountered in the texts makes the annotation applicable only to that particular genre/category. The scheme is incapable of encoding the full range of syntactic complexity encountered in texts of different genres.

In more recent work, Evans and Orăsan (2013) addressed these weaknesses by considering three broad classes of signs: *left subordination boundaries*, *right subordination boundaries* and *coordinators*. The classification scheme was also extended to enable the encoding of links and boundaries between a wider range of syntactic constituents to cover more syntactic phenomena. The current paper presents a method to classify signs of syntactic complexity using the annotated dataset they developed.

3 Annotation Scheme

The annotated signs comprise three conjunctions ([*and*], [*but*], [*or*]), one complementiser ([*that*]), six wh-words ([*what*], [*when*], [*where*], [*which*], [*while*], [*who*]), three punctuation marks ([*,*], [*;*], [*:*]), and 30 compound signs consisting of one of these lexical items immediately preceded by a punctuation mark (e.g. [*, and*]). In this paper, signs of coordination are referred to as *coordinators* whereas signs of subordination are referred to as *subordination boundaries*. In the annotation scheme, the class labels, also called sign tags, are acronyms expressing four types of information:

1. $\{C|SS|ES\}$, the generic function as a coordinator (C), the left boundary of a subordinate constituent (SS), or the right boundary of a subordinate constituent (ES).

2. $\{P|L|I|M|E\}$, the syntactic projection level of the constituent(s): prefix (P), lexical (L), intermediate (I), maximal (M), or extended/clausal (E).

3. $\{A|Adv|N|P|Q|V\}$, the grammatical category of the constituent(s): adjectival (A), adverbial (Adv), nominal (N), prepositional (P), quantificational (Q), and verbal (V).

4. $\{1|2\}$, used to further differentiate sub-classes on the basis of some other label-specific criterion.

The annotation scheme also includes classes which bound interjections, tag questions, and reported speech and a class denoting false signs of syntactic complexity, such as use of the word *that* as a specifier or anaphor.

Collection	Genre	Signs
1. METER corpus	News	12718
2. *www.patient.co.uk*	Healthcare	10796
3. Gutenberg	Literature	11204

Table 1: Characteristics of the annotated dataset.

Signs of syntactic complexity occurring in texts belonging to three categories/genres were annotated in accordance with this scheme[3]. Their characteristics are summarised in Table 1. Absolute and cumulative frequencies of signs and tags reveal a skewed distribution in each genre, e.g. in the news corpus 15 of 40 tags and 11 of 29 signs account for more than 90% of total occurrences.

In the context of information extraction, Evans (2011) showed that automatic syntactic simplification can be performed by annotating input sentences with information on the parts of speech of words and the syntactic functions of coordinators. These annotated sentences can then be simplified according to an iterative algorithm which aggregates several methods to identify specific

[2]Conjoins are the elements linked in coordination.

[3]The annotated dataset and a description of each sign is available at `http://clg.wlv.ac.uk/resources/SignsOfSyntacticComplexity/`

syntactic patterns and then transform the input sentence into several simpler sentences. Each pattern is recognised on the basis of the class assigned to the sign which triggers it and the words surrounding the sign, and is rewritten according to manually created rules.

When a particular syntactic pattern is recognised, a rewriting rule is activated which identifies coordinated structures, the conjoins linked in coordination, and subordinated constituents. Each sign triggers the activation of a simplification rule. The rule applied varies according to the specific class to which the sign belongs.

One advantage of this general approach to syntactic simplification is that it does not depend on syntactic parsing, a process whose reliability depends both on the characteristics of the treebank exploited in training and on the length and complexity of the sentences being processed (McDonald and Nivre, 2011). Another advantage is its flexibility: subsets of rewriting operations can be activated in accordance with user requirements.

4 Tagging Signs of Syntactic Complexity

4.1 Approach

The automatic classification of signs of syntactic complexity is challenging because of the skewed nature of the dataset. As mentioned in Section 2, Evans (2011) proposed a supervised approach to distinguish different types of coordinators in order improve relation extraction from biomedical texts. For each occurrence of a coordinator, a separate training instance was created to describe the surrounding context and then a statistical classifier was built for each coordinator. In that work, experiments were carried out with different classification models such as decision trees, SVM, and naïve Bayes. The best results were obtained by a memory-based learning (MBL) classifier.

In addition to the approach proposed by Evans (2011), we also built and evaluated CRF tagging models (Lafferty et al., 2001; Sutton and McCallum, 2010). These models perform joint inference which can better exploit interactions between different signs present in one sentence, and leads to better performance than is possible when each sign is classified independently. CRF models also achieve state of the art performance in many sequence tagging tasks such as named entity recognition (Tjong Kim Sang and De Meulder, 2003; McCallum and Li, 2003; Settles, 2004), biomedical information extraction (Settles, 2005) or shallow parsing (Sha and Pereira, 2003).

In the annotated dataset, signs of syntactic complexity typically delimit syntactic constituents. Each sign has a tag which reflects the types of constituent it links or bounds. For coordinators, the tag reflects the syntactic category of its conjoins. For subordination boundaries, the tag reflects the syntactic category or type of the bound constituent. This annotation is sign-centric, meaning that the actual extent and type of constituents is not explicitly annotated. To employ a tagging approach, the dataset needs to be converted to a suitable format.

4.2 Tagging Modes

A straightforward way to convert the annotated corpus into a sequence tagging dataset is to consider each sign as a single token chunk whose tag encodes specific information about its syntactic linking or bounding function (Section 2). All the other words are considered as being external to these chunks (tagged as NA). The weakness of this approach is that a baseline predicting the tag NA for every token, providing no useful information, achieves an overall token accuracy greater than 90% because less than 10% of tokens are signs. This can have negative implications for the convergence of the model.

Another mode, inspired by the BIO model adopted in NLP tasks such as named entity recognition (Nadeau and Sekine, 2007) or shallow parsing (Sha and Pereira, 2003), assigns each token the tag of the nearest preceding sign. This amounts to considering the sentence to be split into a set of non-overlapping chunks, each starting with a sign of syntactic complexity. A baseline applying the most common tag (SSEV[4]) to every token achieves an accuracy of 26%, much lower than in the previous setting. The two modes use equivalent information, but in the second mode both signs and words influence the overall tagging of the sentence, which can sometimes lead to different predictions than those made by the first tagging mode. The accuracy of the two modes is compared in Section 5. To have a more informative estimation of performance, only tags assigned to signs are considered for evaluation, while the tags predicted for other tokens are ignored.

[4]Denoting the left boundary of a subordinate clause.

4.3 Tag Sets

As noted in Section 2, the simplification algorithm processes syntactic complexity by iterative application of simplification rules that are specific to signs with particular tags. Given that, when simplifying a specific phenomenon not all tags are necessarily relevant, one research question is whether it is better to use a single CRF model, trained using the complete tagset, or to train a more specialised CRF model instead, using a reduced tagset in which tags irrelevant for the simplification process are combined into a few generic tags. This issue is also investigated in Section 5.

4.4 Feature Sets

The features proposed by Evans (2011) included information about each potential coordinator and its surrounding context (a window of 10 tokens and their POS tags), together with information on the distance of the potential coordinator to other instances in the same sentence and the types of these potential coordinators. This is called the *extended* feature set.

A statistical significance analysis of the extended features showed that most features have very low χ^2 score and that supervised classifiers achieve similar performance when only the features of surrounding tokens are used, i.e. word form and POS tag. This is called the *core* feature set. We investigate whether this finding is observed for the CRF models in Section 5.

5 Evaluation and Discussion

5.1 Setting of the Experiment

Table 1 gives an overview of the size of the annotated corpus described in (Evans and Orăsan, 2013). Sentences from this dataset which contain annotated signs of syntactic complexity were extracted, tokenised and POS-tagged using GATE (Cunningham, 2002). For each genre, sentences were shuffled and split into 10 folds to carry out experiments using cross validation.

Both signs and tags have a skewed distribution. More than 90% of occurrences consist of less than half of the set of tags. A similar observation can be made for the different signs. This makes it difficult to build accurate models for infrequent tags which together comprise less than 10% of occurrences.

An objective of this study is to determine the set of features that are most effective for tagging signs of syntactic complexity. The core feature set is based on word forms and POS tags which are generic features which can be easily and reliably extracted. Evans (2011) uses a more comprehensive set of features. We have employed that system to extract additional features for the annotated signs, the extended set. This also affords an indirect comparison between the classification approach and the sequence labelling approach. Since that system creates a classification instance for each sign independently, in order to use the additional features in a sequence labelling model, an additional unigram CRF template was created for each feature to condition the tag of a sign. As these features are only computed for signs, no templates were used to link the feature values to those of neighbouring tokens. The approach of Evans (2011) was also employed as a baseline (i.e. training supervised classification models which predict a label for each sign independently using the extended set of features) to compare the performance of the CRF model on this dataset. Table 2 shows that the extended feature set (CRF-extended) improves results of the simple tagging on the news genre by 2 points compared to the model using just words and their POS (CRF-core). The table also shows the performance of the baseline approach, when training standard classifiers from Weka (Hall et al., 2009). Regardless of the classifier model used, the baseline approach performs substantially worse than the sequence tagging models. In the following sections all experiments are carried out using CRFs.

	Correct	Accuracy
CRF-extended	10248	80.58%
CRF-core	9979	78.46%
SMO	7213	56.71%
NB	6712	52.78%
J48	6742	53.01%
IB7	6662	52.38%

Table 2: Performance on news corpus using the extended features proposed by Evans (2011)

5.2 Results on the Whole Corpus

Table 3 shows the results achieved for each of the three genres when using two tagging modes, simple and BIO, and two different tag sets, complete and reduced. Results were computed using 10-fold cross-validation. For both news and literature corpora, using the BIO tagging mode leads to better

Genre	tagging	tagset	**P**	**R**	**F1**	Signs	Correct	Incorrect
news	simple	complete	0.7971	0.7846	0.7894	12718	9979	2739
news	BIO	complete	0.8157	0.7991	**0.8053**	12718	10163	2555
literature	simple	complete	0.8414	0.8267	0.8326	11204	9262	1942
literature	BIO	complete	0.8597	**0.8383**	0.8468	11204	9392	1812
healthcare	simple	complete	0.8422	0.8323	**0.8358**	10796	8985	1811
healthcare	BIO	complete	0.8406	0.8244	0.8300	10796	8900	1896
news	simple	reduced	0.8206	0.8161	0.8176	12718	10379	2339
news	BIO	reduced	0.8382	0.8328	**0.8348**	12718	10592	2126
literature	simple	reduced	0.8698	0.8595	0.8639	11204	9630	1574
literature	BIO	reduced	0.8840	**0.8680**	0.8746	11204	9725	1479
healthcare	simple	reduced	0.8636	0.8567	**0.8593**	10796	9249	1547
healthcare	BIO	reduced	0.8602	0.8510	0.8544	10796	9187	1609

Table 3: Overall performance using 10-fold cross validation on the three genres, using two tagging modes (simple and BIO) and two tagsets (complete and reduced)

performance than using simple tagging, while the opposite is true for the health corpus.

One of the objectives of these experiments is to establish whether using a reduced tag set offers performance benefits. When tackling a specific syntactic phenomenon, only a subset of signs and tags may be involved. For example, a set of 11 tags were identified which are relevant for detecting appositions and other noun post-modifiers. The remainder were combined into three coarse grained tags indicating the generic function of the sign as the start (SS) or end (ES) of a subordinated constituent or as coordinator (C) of two constituents. These correspond to the first level used for the class labels in the annotated dataset. Performance achieved with the full and the reduced tag set is listed in Table 3. For all genres and irrespective of tagging mode, using the reduced tag set leads to a performance increase of 2-3 percentiles. A more detailed analysis however reveals that this performance increase is not linked to the relevant 11 original tags, but to the 3 coarse tags.

For example, in the news dataset, the three coarse tags account for 35.84% of all signs. Although the reduced tag set demonstrates a 50% error reduction for two signs (*and*, *or*), the performance for the other signs is largely unchanged. The performance on the 11 tags of interest is also unchanged. This result suggests that using a reduced tag set yields a more informative performance estimation for some specific task because irrelevant tagging errors are not taken into account, but it does not necessarily lead to increased performance

	Test genre		
Train genre	news	healthcare	literature
news	**78.46%**	63.96%	69.98%
healthcare	44.95%	**83.23%**	48.74%
literature	62.59%	58.53%	**82.67%**

(a) Simple tagging mode

	Test genre		
Train genre	news	healthcare	literature
news	**79.91%**	61.29%	71.48%
healthcare	48.75%	**82.44%**	51.95%
literature	64.03%	56.44%	**83.83%**

(b) BIO tagging mode

Table 4: Cross-genre F_1 performance of the tagging models; main diagonal represents performance using 10-fold cross-validation

for the relevant original tags. Therefore there is no real benefit in training multiple tagging models with reduced tag sets.

A relevant issue in the context of text simplification is robustness. To gain insights into the strengths and weaknesses of CRF models we measure the impact on performance when models trained on each genre are applied to the other two genres. In this experiment the complete tag set was used. Table 4a) shows the results using simple tagging, while Table 4b) shows the results when using BIO tagging. In both cases, the main diagonal shows within-genre F_1 performance measured using 10-fold cross-validation; the other entries show cross-genre performance. For all

Genre	Signs	Correct	**Accuracy**
merged	34718	28297	81.51%
merged-bio	34718	28642	82.50%
combined	34718	28226	81.30%
combined-bio	34718	28455	81.96%

Table 5: Joint training performance using 10-fold cross validation: merging data from the three genres leads to better performance than that achieved by the best individual models.

models a considerable performance drop can be observed. The news models are the ones that have the best cross-domain performance, while the healthcare models perform worst. This impact on performance is not unexpected, but rather a proof that the three genres differ from a syntactic perspective. In addition, although all genres have a skewed distribution of signs and tags, the actual rankings differ.

To tackle this issue, a supervised genre classifier can be used to detect the genre of a text to select the best model for the genre, however this approach is limited to genres for which annotated data is available. An alternative approach to minimise the effect of over-fitting is to train models using data from all genres. Table 5 compares these two approaches. In the first two runs (merged), 10-fold cross-validation was performed using stratified sampling; each fold in the merged dataset consists of 3 folds, one from each genre. The last two runs (combined), demonstrate the performance achievable when an oracle selects the correct cross-validated model for each prediction. This represents a performance upper-bound since in practice an actual classifier will be used which would introduce additional errors. The experiment indicates that training a single model on the entire dataset (all three genres) yields better results than using the best models for each genre. Although the differences are not large, merging all available data produces a model with superior tagging performance which should also generalise better to new genres.

5.3 Results on the News Corpus

To better understand the nature of the dataset and the performance of the approach, this section presents more in-depth results for the news genre. Although some differences exist, the other two genres are similar (analysis of them is omitted due to space constraints). Tables 6 and 7 show the CRF model's performance on the news genre using 10-fold cross-validation for the most frequent tags and signs, respectively. In terms of micro-averaged statistics the predictions have a good balance between precision and recall. There is more variance when looking at performance of specific tags or signs. For example, some tags such as SSEV, SSCM, SSMA and ESCM have very good performance ($F_1 > 90\%$); most of these tags mark the start of a constituent (the left boundary). Other tags, despite having comparable frequencies are more difficult to identify and only reach substantially lower levels of performance ($F_1 < 70\%$), e.g. CMN_1, ESEV, ESMP, ESMN, ESMA. Most of these signs mark the right boundary of a constituent, which suggests that identifying the end of a constituent is more difficult than identifying the start. This could be caused by multiple embedded constituents, in which the same sign marks the right boundary of several constituents. In such cases, several tags could be considered correct, but in the annotated dataset only the type of the longest constituent was considered: a sign can only have one tag.

A similar situation occurs when looking at the performance achieved per sign in Table 7. Excellent performance ($F_1 > 95\%$) is noted for the complementiser *that* and *wh-* signs such as *who*, *when* or *which*. Due to the skewed distribution, more than 83% of all errors are linked to the two most frequent signs [,] and [and], which only reach F_1 of 75%.

Table 8 shows the feature templates used to train CRF models in these experiments. To evaluate the impact of each feature template, a simple feature selection methodology was employed: a CRF++ model (Kudo, 2005) was trained on the news corpus using a single template and its performance was ranked and compared with a baseline. For this dataset the baseline was considered using the word form of the current token as the single feature, which achieves 40% accuracy. The best templates, reaching 58% accuracy, used part-of-speech trigrams. When used together, the templates in Table 8 achieve 79.91% accuracy when using the simple tagging mode on the news corpus.

5.4 Extrinsic Evaluation

To determine the extrinsic impact of the errors made by the sign classifier, two rule-based syntactic

	Tag	**P**	**R**	**F**$_1$	Support	Cumulative	True-pos	False-pos	False-neg
1	SSEV	0.9642	0.9298	**0.9467**	3275	26%	3045	113	230
2	CMV$_1$	0.8618	0.8083	0.8342	1111	34%	898	144	213
3	CMN$_1$	0.7381	0.6601	*0.6969*	1059	43%	699	248	360
4	CEV	0.8071	0.7795	0.7931	907	50%	707	169	200
5	SSMN	0.8865	0.8384	**0.8618**	885	57%	742	95	143
6	ESEV	0.6383	0.5631	*0.5984*	586	62%	330	187	256
7	SSCM	0.9659	0.9759	**0.9708**	580	66%	566	20	14
8	SSMA	0.9303	0.9574	**0.943**7	516	70%	494	37	22
9	ESMP	0.5858	0.5611	*0.5732*	499	74%	280	198	219
10	CLN	0.7535	0.6918	0.7214	464	78%	321	105	143
11	SSMP	0.8469	0.8167	0.8315	420	81%	343	62	77
12	ESMN	0.5972	0.6101	0.6036	418	84%	255	172	163
13	SSMV	0.8418	0.8103	0.8258	348	87%	282	53	66
14	ESCM	0.9207	0.9379	**0.9292**	322	90%	302	26	20
15	ESMA	0.6457	0.7049	0.6740	305	92%	215	118	90
avg/total		0.8157	0.7991	0.8053	12718	100%	10163		

Table 6: Per tag performance on the 15 most frequent types of complexity signs in the news corpus using BIO style CRF mode (covering $> 90\%$ of occurrences); the last row shows the weighted average performance (for P, R and F1) and counts (total signs and correct predictions)

	Sign	**P**	**R**	**F**$_1$	Support	Cumulative	Correct	Incorrect
1	,	0.7488	0.7312	0.7377	5443	43%	3980	1463
2	and	0.7778	0.7430	0.7562	2564	63%	1905	659
3	that	0.9608	0.9589	**0.9594**	1313	73%	1259	54
4	who	0.9952	0.9928	**0.9940**	418	77%	415	3
5	,and	0.8089	0.7253	0.7585	324	79%	235	89
6	but	0.8921	0.8658	0.8761	313	82%	271	42
7	when	0.9872	0.9840	**0.9856**	312	84%	307	5
8	or	0.6597	0.5961	*0.6146*	255	86%	152	103
9	,who	1.0000	0.9715	**0.9856**	246	88%	239	7
10	which	1.0000	0.9888	**0.9944**	178	89%	176	2
11	what	0.9867	0.9605	**0.9734**	152	91%	146	6
	Overall	0.8157	0.7991	0.8053	12718	100%	10163	2555

Table 7: Per tag performance on the most frequent signs in the news corpus using BIO style CRF mode (covering $> 90\%$ of occurrences); for each sign micro-averaged P, R and F1, as well as total number of signs and of correct predictions

Template	**Accuracy**	Form	Description
b94	58.13%	%x[0,1]/%x[1,1]	CRF++ Bigram Feature POS-bigram(0,1)
u51	58.29%	%x[-1,1]/%x[0,1]/%x[1,1]	POS-trigram(-1,0,1)
u52	55.20%	%x[0,1]/%x[1,1]/%x[2,1]	POS-trigram(0,1,2)
u47	55.90%	%x[0,1]/%x[1,1]	POS-bigram(0,1)
u32	47.11%	%x[0,0]/%x[1,0]	sign(token and POS)
u00	40.40%	%x[0,0]	sign(token)

Table 8: CRF feature templates which outperform the baseline feature template u00

simplification methods were employed which rely on annotated signs. Each method uses a set of rules to identify certain syntactic structures which are then simplified and was developed using the gold standard annotations. The first method addresses noun post-modifiers, such as appositions, adjectival phrases and relative clauses. When the method is run on the gold standard dataset, 1910 sentences containing noun post-modifiers were identified and simplified. When sign annotations produced using 10-fold cross-validation are used instead, due to classification errors 6.91% fewer sentences are automatically simplified, while the remaining 1778 (93.09%) sentences are still simplified accurately, suggesting that the tagging errors have less impact on this particular method.

The second text simplification method addresses a wider range of syntactic phenomena including coordination. It identifies conjoins and subordinate constituents in complex sentences and re-writes them as sequences of shorter, simpler sentences. When this method is applied on automatic annotations, 22.42% of sentences are no longer simplified by the method, suggesting that the method is more sensitive to tagging errors. These results demonstrate that the automatic sign classifier can usefully be exploited in text simplification applications, especially when addressing specific syntactic phenomena.

6 Related Work

There are two major areas of previous work of relevance to the research described in the current paper. They comprise methods for the automatic classification of signs of syntactic complexity and annotated resources that may be exploited for the development of such approaches.

In closely-related work, van Delden and Gomez (2002) present a system to assign syntactic roles to commas. The classification scheme uses 30 class labels to denote coordinating functions (series commas), boundaries of subordinate constituents (enclosing commas), functions linking and bounding clauses and verb phrases (clausal commas), and bounding direct and indirect speech. There is considerable overlap between their scheme and the dataset used in this paper.

Adopting a two phase approach, van Delden and Gomez (2002) apply 38 finite state automata to part of speech tagged data to derive an initial tagging of commas. After this, information from a tag co-occurrence matrix derived from hand annotated training data is used to improve the initial tagging. The system achieved accuracy of 91-95% in identifying the syntactic function of commas in a collection of encyclopaedia and news articles. This is more accurate than the results reported in the current paper (79-87%), which predicts class labels from a wider selection of classes (44 vs. 30) of a wider variety of signs of syntactic complexity (29 vs. one) in documents from three genres: news, patient healthcare, and literature.

In related work, Maier et al. (2012) proposed the addition of a new annotation layer to disambiguate the role of punctuation in the Penn Treebank. They present a detailed scheme to ensure consistent and reliable manual annotation of commas and semicolons with information to indicate their coordinating function. Compared to the dataset used in this paper, their scheme only encodes coarse-grained information with no discrimination between subclasses of coordinating and non-coordinating functions. The task addressed in the current paper is to tag coordinators and subordination boundaries with more detailed syntactic information about the constituents that they link or bound, the first step in a text simplification application.

7 Conclusions

The decision to tag signs of syntactic complexity with information about pairs of single conjoins or single bound constituents means that in many cases, subordination boundaries and coordinators lack information on the full set of constituents bounded or linked by them. As a result, signs bounding subordinate constituents are often not matched pairs. A second limitation of the scheme is the fact that syntactic complexity not signalled by the signs specified in Section 2 of the current paper cannot be identified. These characteristics of the training data (embedded constituents and missing boundaries) exert a negative influence on tagging right subordination boundaries.

Acknowledgments

The research described in this paper was partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development (FP7-ICT-2011.5.5 FIRST 287607).

References

Rajeev Agarwal and Lois Boggess. 1992. A simple but useful approach to conjunct identification. In *Proceedings of the 30th annual meeting for Computational Linguistics*, pages 15–21, Newark, Delaware. Association for Computational Linguistics.

Hamish Cunningham. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.

Richard Evans and Constantin Orăsan. 2013. Annotating signs of syntactic complexity to support sentence simplification. In Ivan Habernal and Vaclav Matousek, editors, *Text, Speech and Dialogue. Proceedings of the 16th International Conference TSD 2013*, Lecture Notes in Computer Science, Plzen, Czech Republic, September. Springer.

Richard Evans. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing*, 26 (4):371–388.

Laurie Gerber and Eduard H. Hovy. 1998. Improving translation quality by manipulating sentence length. In David Farwell, Laurie Gerber, and Eduard H. Hovy, editors, *AMTA*, volume 1529 of *Lecture Notes in Computer Science*, pages 448–460. Springer.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Marcel Adam Just, Patricia A. Carpenter, Timothy A. Keller, William F. Eddy, and Keith R. Thulborn. 1996. Brain activation modulated by sentence comprehension. *Science*, 274:114–116.

Taku Kudo. 2005. CRF++: Yet another CRF toolkit. *Software available at http://crfpp. sourceforge. net*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann.

Wolfgang Maier, Sandra Kübler, Erhard Hinrichs, and Julia Kriwanek. 2012. Annotating Coordination in the Penn Treebank. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 166–174, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Juan Martos, Sandra Freire, Ana Gonzalez, and David Gil. 2013. D2.2 user preferences: updated report. Technical report. Also available as `http://www.first-asd.eu/D2.2`.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

Ryan T. McDonald and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Thomas C. Rindflesch, Jayant V. Rajan, and Lawrence Hunter. 2000. Extracting molecular binding relationships from biomedical text. In *Proceedings of the sixth conference on Applied natural language processing*, pages 188–195, Seattle, Washington. Association of Computational Linguistics.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics.

Burr Settles. 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics.

Charles Sutton and Andrew McCallum. 2010. An introduction to conditional random fields. *arXiv preprint arXiv:1011.4088*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Masaru Tomita. 1985. *Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems*. Kluwer Academic Publishers, Norwell, MA, USA.

Sebastian van Delden and Fernando Gomez. 2002. Combining finite state automata and a greedy learning algorithm to determine the syntactic roles of commas. In *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '02, pages 293–, Washington, DC, USA. IEEE Computer Society.