Voting Combination of Sentences Splitting Classifiers Applied to Several Types of Texts

Luís Carlos dos Santos Marujo

September 2010

INESC-ID Technical Report 45/2010

Table of Contents

11-910 and 11-712: Research and NLP Lab (Spring/Summer 2010)

VOTING COMBINATION OF SENTENCE SPLITTING CLASSIFIERS APPLIED TO SEVERAL TYPES OF TEXT

Luís Marujo

INESC-ID Lisboa, Instituto Superior Técnico, Lisboa, Portugal
Language Technology Institute, Carnegie Mellon University Pittsburgh, PA 15213, USA

ABSTRACT

This work proposes an approach to sentence splitting as a procedure for syntactic simplification. Based on the voted combination of several classifiers and text sources, it achieves 88.9% precision and 89.5% recall. We noted a strong tendency for the number of recommended sentence splits to increase with the readability of the texts. The inclusion of a passive sentence classifier, achieving 96.2% precision and 97.4% recall, into sentence splitting classifier feature set did not improve its performance.
Experiments on different text types revealed that subtitles need a smaller number of sentence splitting operations than other types of texts, such as Euronews articles. This finding might have implications in the development of text simplification tools aiming at reducing the cognitive load of multimedia tools applied to CALL systems using subtitles.

Index Terms — Text Simplification, Sentence Splitting classifiers, Passive Sentence classifier, Voted classifiers, Subtitling, Euronews, Ted Talks, Scholastic texts, Readability level.

1. Introduction and related work

Content delivery over the Internet, namely news, is produced having in mind a determined target population. Although typical readers are native speakers, if their scholastic background is lower than the 9th grade, then they might struggle to read the news [1]. We note that this fact can be even more prominent for second language learners. As a matter of fact, the English and Portuguese REAP systems [2][1] obtained their reading texts from the web. But upon further analysis, we noted that the amount of documents in the low readability levels that are appropriate for learning purposes is scarce. Therefore, transforming high readability level documents to make them easier for students to read and understand seems appropriate. Simplification techniques on the lexical and syntactic level are relatively recent [3] and require further improvements. Splitting complex sentences into simple ones is one example of a syntax level modification and it is the primary focus of this work. We were also motivated by Gasperin et al. [12] work on text simplification for Brazilian Portuguese that have indicated that sentence splitting is the most frequent syntactic operation used when building a simplified text.
There are two categories of text simplification approaches. The earlier systems were typically rule-based, whereas current state of art systems use corpus-based approaches. The rule-based systems [3][10] contain a set of manually created simplifications rules. Parser structures and reduced number of simplification operations, such as splitting relative clauses, compose the rule based systems. Corpus-based systems [11][12] are focused on learning from aligned corpus which simplification operations are necessary.
Based on the assumption that splitting sentences makes them less complex and thus easier to understand, Petersen and Ostendorf [11] shed some light into the topic after analyzing several features. They concluded that length of the sentence and noun phrases are the most important features. Medero and Ostendorf [13] showed that POS counts, translation counts and definition length, by text mining Wiktionary, are different in the two levels of text used (hard vs. simple). Siddharthan [10] explored syntactic simplification for aphasics. He had the overarching concern of preserving the original structure and cohesion of the document.
Developing a general-purpose sentence splitting classifier can bring improvements to other areas. Parsing long sentences poses several problems as time and memory requirements, e.g. the Stanford Parser's memory usage expands roughly with the square of the sentence length. In statistical machine translation, long sentences cause problems in word alignment, where sentence segmentation helps solving these problems by splitting them and achieving improvements of more than 20% in the BLEU score [4]. Thus, automatic identification of splittable long

sentences might help solve or at least mitigate these issues. But for the purpose of making our objective clear consider, for instance, the next 2 sentences as examples of what we would like to identify. While the first sentence is an example of a non-splittable sentence, the second sentence mirrors a splittable one:

- One time I went to Mexico. - Sentence retrieved from a grade one scholastic text in the training corpus.
- But great allowances should be given to a king, who lives wholly secluded from the rest of the world, and must therefore be altogether unacquainted with the manners and customs that most prevail in other nations : the want of which knowledge will ever produce many prejudices, and a certain narrowness of thinking, from which we, and the politer countries of Europe, are wholly exempted. - Sentence retrieved from a grade eleven scholastic text in the training corpus.

This work explores the construction of a sentence splitting classifier and its applications. It is our hope that this classifier and the following study contribute to advance of the state of art of text simplification. In addition, we hope that this work could be also useful for other knowledge domains as Machine Translation, Text Summarization, etc.
Our work also looks at subtitles in order to enrich them with information about sentence splitting with a target of simplifying them. Subtitled video programs can enhance foreign language learning [5][6] and improve reading skills [5]. But one advantage of using subtitled movies lies in the fact that its use can increase students' motivation [7]. Finally, simplifying subtitles might improve accessibility to entertainment, information, and news, namely subtitled live broadcast news [8][9].

2. DESCRIPTION OF THE STUDY

We set up a study to examine the variability of the percentage of the sentence splitting across several types of text. Determine which one is more suitable for applying text simplification with learning purposes is the goal of this work. We collected text from several sources: news (Euronews), scholastic text (texts from US state exams), and spoken text (Ted talks subtitles containing punctuaction) to attain this goal. For the purpose of this study, we defined 2 classes of syntactic simplification: splittable (S) and non-splittable (N) sentences. The goal of this study is to find which type of text requires less text splitting operations to be able to simplify texts with educational purposes.
Figure 1 shows the high-level division of the main blocks used to perform this study. The unlabeled text (table 2) is split into sentences applying MorphAdorner [18]. We decided to use this sentence splitter toolkit because it supports abbreviations, interjections, number (with/without dots) and provided an easy software integration. Spelling correction was achieved by comparing words in the sentences to entries in CMU-DICT [14]. Spelling errors were corrected by finding them in a list of common mistakes [15]. Finally the Lucene SpellChecker [16] was used to correct the remaining errors. After the preprocessing pass, the features were extracted. Then using the best classifier (voting based and trained using all 3 sources of text) we labeled the sentences from the evaluation set.
We have built a passive sentence classifier motivated by the idea that analyzing voice constructions could boost the sentence splitting classifier performance. Despite the fact that we were also interested in identifying reduced passive verb and sentences containing both passive and active voice, as the table 1 shows, the percentage of reduced passives is only near 0,4%, both passive and active sentences around 1,7%, and passive voice sentences about 1,8%. These small percentages reflect the low frequency of these syntactic constructions. Hence, we included reduced passive verb sentences and sentences containing both passive and active voice as the passive sentences. Additionally, we extracted passive sentences from several webpages explaining the passive voice construction. Including this extra set of sentences was essential because the total of passive voice based sentences in the other 3 corpus collected is below 4% (≈3,9%).
The evaluation is divided in 2 types. First, we measured the percentage of recommend split sentences on the datasets. While the training set and a labeled sample of the test set were labeled and consequently counting the percentage of split is straightforward, we cannot gain a further insight without analyzing a larger dataset. Thus, we applied the sentence splitting classifier to the test set. This leads to the evaluation of the sentence splitting classifier. We evaluated it using 10-fold cross validation on the training set and by direct comparison of the percentage of splits found in the labeled sample.

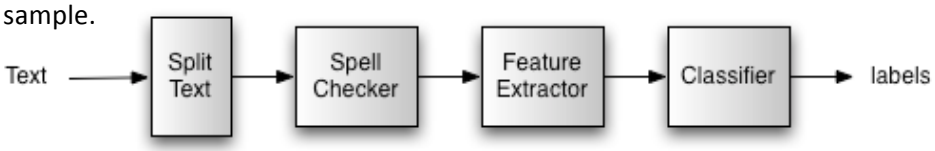

Fig. 1: System architecture

2.1. Gold standards

The gold standard consists of several labeled sentences annotated by the author. We ask a graduate student volunteer (native speaker of English) to label (split/non-split) a random sample of 703 sentences (about 12% of the sentence splitting gold standard). Then we used the Cohen Kappa to analyze the . The results show that the two labelers substantially agree [17]. We obtained a Cohen Kappa of 0.606 (SE = 0.035, p ≈ 0). Then we repeat the same procedure to evaluate the passive sentences gold standard. We extracted a random sample of 700 sentences (about 10% of the passive sentence gold standard). We achieved an excellent agreement [17], i.e. Cohen Kappa of 0.951 (SE = 0.0075, p≈ 0).

Formatting and labeling several texts in a sentence per line is a rather time consuming task. Therefore, we built a small command line annotation editor to speed up this process. The sentence boundaries are detected using MorphAdorner.

Source	#Docs	# Active	#Passive	#Reduced	#Pas. and Active	#Rec. Sent. Splits	#Sentence
Euronews	208	1799	68	14	89	1024 (52%)	1970 (28%)
Educational texts	35	1959	22	8	8	738 (37%)	1997 (28%)
Ted Talks	11	1893	14	1	5	900 (47%)	1915 (27%)
Passive sentences*	–	–	1099	–	–	-	1099 (16%)
Total		**5652 (81%)**	**1328 (19%)**				**6981**

Table 1: Training Datasets manually labeled (gold standard)

* – dataset included to improve passive sentences classifier performance.

Source	# Docs	# Sentences	Sample (# Sentences)
Euronews	567	4033	336 (8,331%)
Scholastic texts	Gr. 01 : 8	204	34 (16,667%)
	Gr. 02 : 2	202	124 (61,386%)
	Gr. 03 : 42	3661	57 (1,557%)
	Gr. 04 : 43	3820	161 (4,215%)
	Gr. 05 : 40	4515	216 (4,784%)
	Gr. 06 : 34	4057	102 (2,514%)
	Gr. 07 : 36	4817	184 (3,820%)
	Gr. 08 : 40	4750	91 (1,916%)
	Gr. 09 : 13	1285	15 (1,167%)
	Gr. 10 : 18	3308	22 (0.665%
	Gr. 11 : 4	486	36 (7,407%)
	Gr. 12 : 11	365	26 (7,123%)
	Total : 566	31470	1067 (3,339%)
Ted Talks	330	53601	787 (1,468%)

Table 2: Evaluation Datasets (do not include labels except the sample)

Scholastic texts entries include the number of documents and sentences per grade level to complement Fig. 3.

2.2. Sentence Splitting classifier

We followed and extended the sentence splitting classifier approach (SSC) proposed by Gasperin et al. [12] for Brazilian Portuguese. The feature extraction process is the first step of the classification process. It extracts the following frequency based features:

1. lexical features – number of characters, words, proper names;
2. syntactic features – number of occurrences of each POS tag, described in Marcus et al. [19];

3. syntactic structural features - number and average size of Penn tree-bank style phrase structures: NP (Noun Phrases), VP (Verb Phrases), SBAR (Subordinate Clause), ADJP (Adjective Phrase), ADVP (Adverb Phrase), PP (Prepositional Phrase);
4. cue phrases features - number of cue phrases;
5. cue phrases position ratio features (CPPR) - position of cue phrases in the sentences;
6. typed dependency features (TD)– number of occurrences of each dependency parsing labels, described in Marneffe et al.[34][35];
7. typed dependency position ratio features (TDPR) – position of typed dependency features;
8. Passive classifier label;

The cue phrase list was initially based on the best performing cue phrases for Brazilian Portuguese: *accordingly, actually, additionally, afterwards, although, as a matter of fact, before, but, clearly, consequently, even so, hence, in addition, in case, in comparation, in contrast, in fact, in general, namely, nevertheless, nor, now, on the other hand, or, posteriorly, regarding, similarly, since, so that, speaking of, that, then, thereby, this way, thus, to, until, which, with, yet.*

During the annotation process we observed that the following phrases could also be relevant features: *and, as, despite, however, on the other hand, said, say, says, where, who, while,* punctuation marks (“,”, “;”, “?” ,“!”, “.”), and anaphoric words, such as pronouns: I, you, he, she, it, we, they.

While for extracting Penn Treebank phrase structures we used the Stanford Parser [20], POS tags are captured by the Stanford POS tagger [21] . There are two reasons that justify such design decision. The first reason concerns the higher accuracy of the POS tagger when compared to the parser. The other reason is correlated with the recurrent trade-off accuracy vs. computational costs, which, under some circumstances, can play an important role leading to the exclusion of the parser based features. By counting the number of occurrences of each tag in the Penn Treebank English POS tag set occurring in each sentence, i.e. counting the number of verb tags, we are also implicitly detecting passive and reduced passive voice. MorphAdorner was the toolkit chosen to identify name entities (proper names).

The cue phrase position ratio feature was not part of the original list of features in [12] (basic features). The conceptual idea behind its inclusion was to consider the position of the cue phrases in the sentences.

Defining the position, $p(w_f \mid s_i)$, as the number of words before the word or phrase (w_f) under consideration in a sentence s_i. Given that sentence length is a random variable, it is necessary to normalize the position by sentence length, i.e., creating a position ratio. In addition, a cue phrase can appear several times in a sentence. Instead of considering all positions, we decided to choose their centroid, i.e. their average position in the sentence. Hence, for each word/phrase in the cue list w_i, the observed position or cue phrase position ratio o_c in the sentence is comp

$$o_c(w_i) = \frac{1}{L}\sum_{i=0}^{I}\frac{1}{i}p(w_i|s_f) \tag{1}$$

Considering the following sentence found in the Euronews corpus: “The house has an indoor swimming pool and even an indoor garden and Milenkovic says that the savings in electricity mean that his investment will have paid for itself in just four years.”, we have the following cue phrase position ratio features:

- and - $\frac{\frac{8+13}{2}}{32} = \frac{21}{64}$
- that - $\frac{\frac{16+22}{2}}{32} = \frac{19}{32}$
- says - $\frac{15}{32}$

(2)

Preliminarily, we retrieved the typed dependency features to identify passive sentences, because the “auxpassive” label usually appears in the extracted dependency list of a passive sentence. However, the number of occurrences

and ratio position of the remaining 58 dependency tags also provide relevant information. Thus, we included them as features.

We also use the passive classifier, described in section 3.2, labels as feature.

After running the feature extraction process over the training sets, we created several models using several classifiers available in the Weka tookit [22]. Table 3 shows the results. We trained SMO[23], Bayesian Logistic Regression [25], Simple Logistic [26], Voted Perceptron [27], Simple Cart [28], and C4.5 [29].

2.3. Passive sentence classifier

The passive classifier shares the same architecture of the sentence splitting classifier. The main difference concerns the clue phrases used. The list of clue phrases has modals ("can", "could", "may", "might", "must", "ought to", "shall", "should", "will", "would"). We considered also auxiliary verbs ("be", "have", "do", "become", "appear", "feel", "get", "remain", "seem") in present, past and gerund tense. The list of auxiliary verbs was based on the list used by Igo and Rillof [33] to identify passive and reduced passive verb phrases. In addition, linking words such as "by", "who", "that", and "which" are also part of the list based on our empiric analysis.

3. RESULTS

Table 4 shows the results obtained for the sentence splitting classifiers using each training dataset and all features. Table 3 contains its corresponding confusion matrix. Then, in table 5, we explored different combinations of features and its influence on the classifiers performance. Table 6,8 and 10 emerged as complementary tables describing the confusion matrix of the classifiers whose performance values are presented in their corresponding previous table. We also tested the passive classifier performance in table 7 (considering each training dataset and all features) and table 8 (using all features and training data available).

The voted combination of several classifiers yielded the best performance. The following Figures 2 and 3 were generated using the voted based classifier. We already knew the global percentage of recommended splits for scholastic texts from Figure 3. However, we decided to increase the level of detail by analyzing the distribution of recommend percentage of splits per school grade level, as Figure 3 shows.

It is important to pay attention to the fact that in both Figure 2 and 3 we verified that the percentage differences are statistically significant ($p \approx 0$) under a Chi-Square Test.

Feature Set	Class	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Ted Talks	P. S.	177	183	175	133	126	199	82
	A. S.	723	717	725	767	774	701	418
	P. N.	129	129	128	184	199	187	93
	A. N.	885	885	886	830	815	827	959
	U.	0	0	0	0	0	0	445 (22.283%)
Educational texts	P. S.	197	252	249	183	224	227	64
	A. S.	544	489	492	558	517	514	617
	P. N.	174	136	127	203	197	197	81
	A. N.	1082	1120	1129	1053	1059	1059	755
	U.	0	0	0	0	0	0	397 (20.742%)
Euronews	P. S.	189	184	186	125	174	209	67
	A. S.	925	830	828	889	840	805	724
	P. N.	162	167	146	225	203	200	101
	A. N.	794	789	810	731	753	756	658
	U.	0	0	0	0	0	0	420 (21.320%)

Table 3: Confusion matrix of evaluation of the sentence splitting classifier (see Table 4).
Legend: P.S. – Predicted splits , A.S. – Actual splits , P.N. – Predicted non-splits , U. – Unclassified Instances
Note: Voted combination generates unclassified instances when the product of probabilities underflow.

Feature Set	Class	Measure	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Ted Talks	Split	P	0.849	0.848	0.850	0.807	0.795	0.789	0.884
		R	0.803	0.797	0.806	0.852	0.860	0.779	0.906
		F1	0.825	0.821	0.827	0.829	0.826	0.784	0.895
	Non-split	P	0.833	0.829	0.835	0.862	0.866	0.806	0.922
		R	0.873	0.873	0.874	0.819	0.804	0.816	0.903
		F1	0.853	0.850	0.854	0.840	0.834	0.811	0.912
Educational texts	Split	P	0.758	0.782	0.795	0.733	0.724	0.723	0.818
		R	0.734	0.660	0.664	0.753	0.698	0.694	0.836
		F1	0.746	0.716	0.724	0.743	0.711	0.708	0.827
	Non-split	P	0.846	0.816	0.819	0.852	0.825	0.823	0.921
		R	0.861	0.892	0.899	0.838	0.843	0.843	0.912
		F1	0.854	0.852	0.857	0.845	0.834	0.833	0.916
Euronews	Split	P	0.836	0.832	0.850	0.798	0.805	0.801	0.878
		R	0.814	0.819	0.817	0.877	0.828	0.794	0.915
		F1	0.825	0.825	0.833	0.836	0.817	0.797	0.896
	Non-split	P	0.808	0.811	0.813	0.854	0.812	0.783	0.908
		R	0.831	0.825	0.847	0.765	0.788	0.791	0.867
		F1	0.819	0.818	0.830	0.807	0.800	0.787	0.887

Table 4: Evaluation of the sentence splitting classifier using 10-fold cross-validation and all features.

Feature Set	Class	Measure	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Basic	Split	P	0.826	0.802	0.823	0.756	0.782	0.775	0.838
		R	0.802	0.757	0.760	0.846	0.791	0.776	0.876
		F1	0.813	0.779	0.790	0.799	0.786	0.776	0.857
	Non- split	P	0.841	0.809	0.814	0.860	0.826	0.816	0.896
		R	0.861	0.846	0.866	0.776	0.818	0.815	0.865
		F1	0.851	0.827	0.839	0.816	0.822	0.815	0.880
CPPR	Split	P	0.786	0.784	0.744	0.760	0.722	0.755	0.861
		R	0.654	0.659	0.612	0.698	0.704	0.669	0.885
		F1	0.714	0.716	0.671	0.727	0.713	0.709	0.872
	Non- split	P	0.750	0.752	0.721	0.767	0.761	0.751	0.906
		R	0.853	0.850	0.827	0.819	0.777	0.821	0.887
		F1	0.798	0.798	0.771	0.792	0.769	0.784	0.896
CPPR + TD	Split	P	0.832	0.776	0.823	0.811	0.771	0.778	0.852
		R	0.778	0.786	0.744	0.838	0.771	0.744	0.845
		F1	0.804	0.781	0.781	0.824	0.771	0.761	0.848
	Non- split	P	0.826	0.822	0.805	0.863	0.812	0.797	0.877
		R	0.870	0.813	0.868	0.839	0.811	0.826	0.884
		F1	0.848	0.818	0.835	0.851	0.811	0.811	0.881

Table 5 – Part 1 of 5: Evaluation of sentence splitting classifier using 10-fold cross-validation and all 3-training corpus.

Feature Set	Class	Measure	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
CPPR + TDPR	Split	P	0.830	0.812	0.813	0.764	0.778	0.770	0.857
		R	0.730	0.731	0.716	0.829	0.797	0.756	0.862
		F1	0.777	0.770	0.762	0.795	0.788	0.763	0.860
	Non- split	P	0.798	0.796	0.787	0.848	0.830	0.802	0.892
		R	0.877	0.861	0.865	0.789	0.813	0.814	0.888
		F1	0.835	0.827	0.824	0.818	0.821	0.808	0.890
CPPR + Passive	Split	P	0.830	0.813	0.813	0.763	0.778	0.766	0.851
		R	0.730	0.728	0.716	0.824	0.797	0.754	0.838
		F1	0.776	0.768	0.762	0.792	0.788	0.760	0.844
	Non- split	P	0.798	0.794	0.787	0.845	0.830	0.800	0.875
		R	0.877	0.862	0.865	0.790	0.813	0.811	0.885
		F1	0.835	0.827	0.824	0.816	0.821	0.805	0.880
CPPR + TD + TDPR	Split	P	0.830	0.805	0.821	0.811	0.791	0.777	0.852
		R	0.773	0.762	0.738	0.838	0.805	0.754	0.845
		F1	0.800	0.783	0.777	0.824	0.791	0.766	0.848
	Non- split	P	0.823	0.812	0.801	0.863	0.837	0.803	0.877
		R	0.869	0.848	0.867	0.839	0.825	0.822	0.884
		F1	0.846	0.830	0.833	0.851	0.831	0.812	0.881
CPPR + TD + TDPR + Passive	Split	P	0.830	0.819	0.821	0.767	0.791	0.775	0.857
		R	0.773	0.767	0.738	0.832	0.805	0.756	0.862
		F1	0.800	0.792	0.777	0.798	0.798	0.765	0.860
	Non- split	P	0.823	0.818	0.801	0.851	0.837	0.803	0.892
		R	0.869	0.860	0.867	0.792	0.825	0.820	0.888
		F1	0.846	0.838	0.833	0.821	0.831	0.811	0.890
TD	Split	P	0.824	0.821	0.816	0.787	0.767	0.764	0.830
		R	0.776	0.762	0.735	0.818	0.778	0.744	0.821
		F1	0.799	0.791	0.773	0.802	0.773	0.754	0.825
	Non- split	P	0.824	0.815	0.798	0.846	0.815	0.794	0.856
		R	0.864	0.863	0.863	0.818	0.806	0.811	0.864
		F1	0.843	0.839	0.829	0.831	0.811	0.802	0.860
TD + TDPR	Split	P	0.824	0.814	0.819	0.763	0.781	0.761	0.845
		R	0.773	0.755	0.739	0.833	0.807	0.748	0.845
		F1	0.797	0.784	0.777	0.796	0.793	0.754	0.845
	Non- split	P	0.822	0.810	0.801	0.851	0.836	0.795	0.875
		R	0.864	0.858	0.865	0.787	0.813	0.807	0.875
		F1	0.843	0.833	0.832	0.818	0.825	0.801	0.875

Table 5 – Part 2 of 5: Evaluation of sentence splitting classifier using 10-fold cross-validation and all 3-training corpus.

Feature Set	Class	Measure	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
TD + Passive	Split	P	0.820	0.805	0.816	0.787	0.766	0.763	0.828
		R	0.779	0.769	0.735	0.816	0.786	0.745	0.821
		F1	0.799	0.787	0.773	0.801	0.776	0.754	0.825
	Non- split	P	0.825	0.817	0.798	0.844	0.820	0.794	0.856
		R	0.860	0.847	0.863	0.819	0.803	0.810	0.862
		F1	0.842	0.832	0.829	0.831	0.811	0.802	0.859
TDPR	Split	P	0.817	0.804	0.812	0.754	0.773	0.753	0.826
		R	0.722	0.730	0.713	0.826	0.792	0.753	0.817
		F1	0.766	0.765	0.759	0.788	0.783	0.753	0.822
	Non- split	P	0.791	0.793	0.785	0.844	0.826	0.797	0.858
		R	0.867	0.854	0.864	0.779	0.808	0.797	0.865
		F1	0.827	0.822	0.823	0.810	0.817	0.797	0.861
TDPR + Passive	Split	P	0.819	0.811	0.812	0.754	0.773	0.757	0.828
		R	0.724	0.730	0.713	0.826	0.791	0.760	0.817
		F1	0.768	0.768	0.759	0.788	0.782	0.758	0.823
	Non- split	P	0.792	0.795	0.785	0.844	0.825	0.802	0.858
		R	0.868	0.860	0.864	0.779	0.809	0.799	0.868
		F1	0.829	0.826	0.823	0.810	0.817	0.800	0.863
Passive	Split	P	0.549	0.549	0.549	0.549	0.549	0.549	0.549
		R	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.708	0.708	0.708	0.708	0.708	0.708	0.708
	Non- split	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		R	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		F1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Basic + CPPR	Split	P	0.829	0.81	0.834	0.758	0.791	0.769	0.866
		R	0.787	0.774	0.768	0.842	0.807	0.771	0.877
		F1	0.808	0.791	0.799	0.798	0.799	0.770	0.872
	Non-split	P	0.832	0.820	0.820	0.857	0.838	0.811	0.906
		R	0.866	0.851	0.874	0.779	0.824	0.810	0.887
		F1	0.849	0.835	0.846	0.816	0.831	0.810	0.896
Basic + TD	Split	P	0.843	0.820	0.826	0.764	0.792	0.774	0.866
		R	0.825	0.766	0.767	0.855	0.808	0.773	0.901
		F1	0.834	0.792	0.796	0.807	0.800	0.774	0.883
	Non-split	P	0.859	0.818	0.819	0.868	0.839	0.814	0.918
		R	0.874	0.861	0.867	0.782	0.826	0.814	0.888
		F1	0.866	0.839	0.843	0.823	0.833	0.814	0.903

Table 5 – Part 3 of 5: Evaluation of sentence splitting classifier using 10-fold cross-validation and all 3-training corpus.

Feature Set	Class	Measure	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Basic + TDPR	Split	P	0.838	0.826	0.829	0.772	0.792	0.769	0.856
		R	0.809	0.751	0.763	0.834	0.808	0.763	0.878
		F1	0.824	0.787	0.795	0.802	0.800	0.766	0.867
	Non-split	P	0.847	0.810	0.817	0.854	0.839	0.806	0.904
		R	0.872	0.870	0.870	0.798	0.825	0.811	0.886
		F1	0.859	0.839	0.843	0.825	0.832	0.809	0.895
Basic + Passive	Split	P	0.839	0.826	0.829	0.769	0.792	0.771	0.855
		R	0.811	0.759	0.763	0.833	0.808	0.761	0.882
		F1	0.825	0.791	0.795	0.800	0.800	0.766	0.868
	Non-split	P	0.849	0.814	0.817	0.852	0.839	0.805	0.907
		R	0.872	0.868	0.870	0.794	0.825	0.814	0.886
		F1	0.860	0.840	0.843	0.822	0.832	0.810	0.896
Basic + CPPR + TD	Split	P	0.841	0.812	0.832	0.773	0.793	0.779	0.872
		R	0.814	0.761	0.763	0.855	0.820	0.777	0.905
		F1	0.827	0.785	0.796	0.812	0.806	0.778	0.888
	Non-split	P	0.851	0.813	0.818	0.869	0.848	0.817	0.924
		R	0.873	0.855	0.874	0.793	0.824	0.819	0.896
		F1	0.862	0.833	0.845	0.829	0.836	0.818	0.910
Basic + CPPR + TDPR	Split	P	0.844	0.827	0.825	0.774	0.791	0.774	0.870
		R	0.791	0.755	0.765	0.838	0.798	0.768	0.889
		F1	0.817	0.790	0.794	0.804	0.795	0.771	0.879
	Non-split	P	0.837	0.812	0.817	0.857	0.832	0.810	0.912
		R	0.879	0.870	0.866	0.798	0.827	0.815	0.896
		F1	0.857	0.840	0.841	0.826	0.830	0.813	0.904
Basic + CPPR + Passive	Split	P	0.835	0.818	0.833	0.755	0.793	0.784	0.860
		R	0.791	0.760	0.764	0.847	0.792	0.777	0.885
		F1	0.812	0.788	0.797	0.798	0.793	0.781	0.872
	Non-split	P	0.835	0.814	0.818	0.860	0.829	0.818	0.907
		R	0.871	0.861	0.874	0.774	0.830	0.824	0.886
		F1	0.853	0.837	0.845	0.815	0.830	0.821	0.896
Basic + TD + TDPR	Split	P	0.845	0.831	0.825	0.774	0.791	0.774	0.870
		R	0.816	0.763	0.765	0.838	0.798	0.768	0.889
		F1	0.830	0.796	0.794	0.804	0.795	0.771	0.879
	Non-split	P	0.853	0.817	0.817	0.857	0.832	0.810	0.912
		R	0.877	0.873	0.866	0.798	0.827	0.815	0.896
		F1	0.865	0.844	0.841	0.826	0.830	0.813	0.904

Table 5 – Part 4 of 5: Evaluation of sentence splitting classifier using 10-fold cross-validation and all 3-training corpus.

Feature Set	Class	Measure	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Basic + TD + Passive	Split	P	0.841	0.820	0.826	0.774	0.792	0.774	0.864
		R	0.823	0.765	0.767	0.852	0.808	0.773	0.901
		F1	0.832	0.791	0.796	0.811	0.800	0.773	0.882
	Non-split	P	0.857	0.816	0.819	0.867	0.839	0.813	0.918
		R	0.872	0.862	0.867	0.796	0.826	0.815	0.887
		F1	0.864	0.838	0.843	0.830	0.833	0.814	0.902
Basic + TDPR + Passive	Split	P	0.839	0.826	0.829	0.769	0.792	0.771	0.855
		R	0.811	0.759	0.763	0.833	0.808	0.761	0.882
		F1	0.825	0.791	0.795	0.800	0.800	0.766	0.868
	Non-split	P	0.849	0.814	0.817	0.852	0.839	0.805	0.907
		R	0.872	0.868	0.870	0.794	0.825	0.814	0.886
		F1	0.860	0.840	0.843	0.822	0.832	0.810	0.896
Basic + CPPR + TD + TDPR	Split	P	0.841	0.820	0.830	0.794	0.793	0.776	**0.889**
		R	0.798	0.759	0.766	0.833	0.806	0.774	**0.895**
		F1	0.819	0.788	0.797	0.813	0.799	0.775	**0.892**
	Non-split	P	0.840	0.813	0.819	0.857	0.838	0.814	**0.922**
		R	0.876	0.863	0.871	0.822	0.827	0.816	**0.918**
		F1	0.858	0.837	0.844	0.839	0.832	0.815	**0.920**
All	Split	P	0.832	0.829	0.835	0.779	0.796	0.787	0.881
		R	0.795	0.789	0.765	0.840	0.800	0.769	0.888
		F1	0.813	0.808	0.798	0.808	0.798	0.778	0.884
	Non-split	P	0.837	0.833	0.819	0.859	0.835	0.813	0.916
		R	0.868	0.866	0.876	0.804	0.831	0.829	0.911
		F1	0.853	0.849	0.846	0.831	0.833	0.821	0.913

Table 5 – Part 5 of 5: Evaluation of sentence splitting classifier using 10-fold cross-validation and all 3-training corpus.

Feature Set	Class	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Basic	P. S.	448	646	637	408	555	594	273
	A. S.	2778	2009	2018	2247	2100	2061	1921
	P. N.	525	497	433	724	586	597	37
	A. N.	2130	2729	2793	2502	2640	2629	2364
	U.	0	0	0	0	0	0	953 (16.205%)
CPPR	P. S.	919	905	1031	803	786	880	243
	A. S.	1736	1750	1624	1852	1869	1775	1861
	P. N.	473	483	558	585	720	577	301
	A. N.	2753	2743	2668	2641	2506	2649	2355
	U.	0	0	0	0	0	0	1121 (19.061%)

Table 6 – part 1 of 3: Confusion matrix of evaluation of the sentence splitting classifier (see Table 5).
Legend:

- P.S. – Predicted splits
- A.S. – Actual splits
- P.N. – Predicted non-splits
- U. – Unclassified Instances

Note: Voted combination generates unclassified instances when the product of probabilities underflow.

Feature Set	Class	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
CPPR + TD	P. S.	590	567	680	430	607	680	352
	A. S.	2065	2088	1975	2225	2048	1975	1912
	P. N.	418	603	426	520	610	562	331
	A. N.	2808	2623	2800	2706	2616	2664	2521
	U.	0	0	0	0	0	0	765 (13.008%)
CPPR + TDPR	P. S.	718	713	754	680	538	648	332
	A. S.	1937	1942	1901	2546	2117	2007	1780
	P. N.	397	449	436	455	604	600	318
	A. N.	2829	2777	2790	2200	2622	2626	2395
	U.	0	0	0	0	0	0	1056 (17.956%)
CPPR + Passive	P. S.	718	721	754	468	538	652	346
	A. S.	1937	1934	1901	2187	2117	2003	1776
	P. N.	398	446	436	678	604	611	312
	A. N.	2828	2780	2790	2548	2622	2615	2396
	U.	0	0	0	0	0	0	1054 (17.921%)
CPPR + TD + TDPR	P. S.	604	632	696	467	518	652	292
	A. S.	2051	2023	1959	2188	2137	2003	1823
	P. N.	421	490	428	550	564	575	303
	A. N.	2805	2736	2798	2676	2662	2651	2408
	U.	0	0	0	0	0	0	1055 (17.940%)
CPPR + TD + TDPR + Passive	P. S.	604	619	696	451	519	649	301
	A. S.	2051	2036	1959	2204	2136	2006	1819
	P. N.	421	451	428	663	564	582	315
	A. N.	2805	2775	2798	2563	2662	2644	2413
	U.	0	0	0	0	0	0	1033 (17.565%)
TD	P. S.	596	631	704	482	589	680	429
	A. S.	2059	2024	1951	2173	2066	1975	1961
	P. N.	439	441	441	588	626	611	2557
	A. N.	2787	2785	2785	2638	2600	2615	401
	U.	0	0	0	0	0	0	533 (9.063%)
TD + TDPR	P. S.	604	650	692	444	513	670	343
	A. S.	2051	2005	1963	2211	2142	1985	1876
	P. N.	438	457	435	688	602	624	344
	A. N.	2788	2769	2791	2538	2624	2602	2412
	U.	0	0	0	0	0	0	906 (15.406%)
TD + Passive	P. S.	588	613	704	489	569	678	428
	A. S.	2067	2042	1951	2166	2086	1977	1967
	P. N.	453	494	441	585	636	613	409
	A. N.	2773	2732	2785	2641	2590	2613	2551
	U.	0	0	0	0	0	0	526 (8.944%)
TDPR	P. S.	739	718	763	463	551	656	404
	A. S.	1916	1937	1892	2192	2104	1999	1804
	P. N.	2796	472	438	714	618	654	379
	A. N.	430	2754	2788	2512	2608	2572	2435
	U.	0	0	0	0	0	0	859 (14.606%)
TDPR + Passive	P. S.	734	717	763	468	554	638	405
	A. S.	1921	1938	1892	2187	2101	2017	1805
	P. N.	425	453	438	690	617	648	374
	A. N.	2801	2773	2788	2536	2609	2578	2454
	U.	0	0	0	0	0	0	843 (14.334%)
Passive	P. S.	0	0	0	0	0	0	0
	A. S.	2655	2655	2655	2655	2655	2655	2655
	P. N.	0	0	0	0	0	0	0
	A. N.	3226	3226	3226	3226	3226	3226	3226
	U.	0	0	0	0	0	0	0 (0.000%)
Basic + CPPR	P. S.	431	601	617	420	513	609	243
	A. S.	2795	2054	2038	2235	2142	2046	1861
	P. N.	565	482	407	713	567	614	301
	A. N.	2090	2744	2819	2513	2659	2612	2355
	U.	0	0	0	0	0	0	1121 (19.061%)

Table 6 – part 2 of 3: Confusion matrix of evaluation of the sentence splitting classifier (see Table 5).

Feature Set	Class	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Basic + TD	P. S.	464	620	384	510	618	602	212
	A. S.	2191	2035	2271	2145	2037	2053	1920
	P. N.	407	448	703	562	428	600	298
	A. N.	2819	2778	2523	2664	2798	2626	2371
	U.	0	0	0	0	0	0	1080 (18.364%)
Basic + TDPR	P. S.	506	660	628	441	510	628	256
	A. S.	2149	1995	2027	2214	2145	2027	1849
	P. N.	414	419	418	653	565	610	311
	A. N.	2812	2807	2808	2573	2661	2616	2417
	U.	0	0	0	0	0	0	1048 (17.820%)
Basic + Passive	P. S.	502	639	628	444	510	635	247
	A. S.	2153	2016	2027	2211	2145	2020	1840
	P. N.	413	426	418	664	565	2627	312
	A. N.	2813	2800	2808	2562	2661	599	2416
	U.	0	0	0	0	0	0	1066 (18.12%)
Basic + CPPR + TD	P. S.	409	635	628	420	478	593	197
	A. S.	2817	2020	2027	2235	2177	2062	1977
	P. N.	493	469	2818	713	567	585	276
	A. N.	2162	2757	408	2513	2659	2641	2381
	U.	0	0	0	0	0	0	1150 (19.555%)
Basic + CPPR + TDPR	P. S.	389	650	625	537	537	615	232
	A. S.	2837	2005	2030	2118	2118	2040	1857
	P. N.	554	418	431	558	558	596	278
	A. N.	2101	2808	2795	2668	2668	2630	2406
	U.	0	0	0	0	0	0	1108 (18.840%)
Basic + CPPR + Passive	P. S.	556	636	407	626	552	591	332
	A. S.	2099	2019	2248	2029	2103	2064	1780
	P. N.	415	448	728	408	548	567	318
	A. N.	2811	2778	2498	2818	2678	2679	2395
	U.	0	0	0	0	0	0	1056 (17.956%)
Basic + TD + TDPR	P. S.	489	630	625	431	537	615	232
	A. S.	2166	2025	2030	2224	2118	2040	1857
	P. N.	397	411	431	651	558	596	278
	A. N.	2829	2815	2795	2575	2668	2630	2406
	U.	0	0	0	0	0	0	1108 (18.840%)
Basic + TD + Passive	P. S.	470	625	618	376	510	604	210
	A. S.	2185	2030	2037	2279	2145	2051	1913
	P. N.	413	446	428	705	562	598	302
	A. N.	2813	2780	2798	2521	2664	2628	2366
	U.	0	0	0	0	0	0	1090 (18.534%)
Basic + TDPR + Passive	P. S.	502	639	628	444	510	635	247
	A. S.	2153	2016	2027	2211	2145	2020	1840
	P. N.	413	426	418	664	565	599	312
	A. N.	2813	2800	2808	2562	2661	2627	2416
	U.	0	0	0	0	0	0	1066 (18.126%)
Basic + CPPR + TD + TDPR	P. S.	401	641	621	621	516	600	199
	A. S.	2825	2014	2034	2034	2139	2055	1688
	P. N.	537	442	417	417	558	592	211
	A. N.	2118	2782	2809	2807	2668	2634	2358
	U.	0	0	0	0	0	0	1425 (24.231%)
All	P. S.	425	561	625	424	530	614	220
	A. S.	2801	2094	2030	2231	2125	2041	1741
	P. N.	544	432	401	633	545	553	235
	A. N.	2111	2794	2825	2593	261	2673	2392
	U.	0	0	0	0	0	0	1293 (21.986%)

Table 6 – part 3 of 3: Confusion matrix of evaluation of the sentence splitting classifier (see Table 5).

Feature Set	Class	Measure	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Ted Talks	Active	P	0.995	0.993	0.994	0.993	0.994	0.995	0.995
		R	0.991	1.000	0.998	0.999	0.999	0.999	1.000
		F1	0.993	0.996	0.996	0.996	0.997	0.997	0.998
	Passive	P	0.217	0.000	0.429	0.000	0.750	0.833	0.000
		R	0.357	0.000	0.214	0.000	0.214	0.357	0.000
		F1	0.270	0.000	0.286	0.000	0.333	0.500	0.000
Educational texts	Active	P	0.995	0.989	0.991	0.989	0.994	0.992	0.995
		R	0.991	0.999	0.998	0.999	1.000	0.997	0.999
		F1	0.993	0.994	0.995	0.994	0.997	0.995	0.997
	Passive	P	0.217	0.000	0.571	0.000	1.000	0.545	0.000
		R	0.357	0.000	0.182	0.000	0.455	0.273	0.000
		F1	0.270	0.000	0.276	0.000	0.625	0.364	0.000
Euronews	Active	P	0.981	0.955	0.966	0.957	0.965	0.974	0.984
		R	0.971	0.999	0.994	0.995	0.989	0.975	0.998
		F1	0.971	0.977	0.980	0.976	0.977	0.974	0.991
	Passive	P	0.491	0.000	0.657	0.357	0.524	0.453	0.556
		R	0.596	0.000	0.258	0.056	0.247	0.438	0.147
		F1	0.538	0.000	0.371	0.097	0.336	0.446	0.233
Passive sentences	Active	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		R	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		F1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Passive	P	0.999	0.999	0.999	0.999	0.999	0.999	0.999
		R	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		F1	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 7 - Evaluation of Passive sentence classifier using 10-fold cross-validation and all features.

Feature Set	Class	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Ted Talks	P. P.	9	22	11	14	11	9	9
	A. P.	5	0	3	0	3	5	0
	P. A.	18	1	4	1	1	1	0
	A. A.	1881	1973	1895	1898	1898	1898	1878
	U.	0	0	0	0	0	0	26 (1.359%)
Educational texts	P. P.	9	14	4	22	21	16	10
	A. P.	5	0	18	0	1	6	0
	P. A.	18	0	3	1	2	5	1
	A. A.	1881	1899	1971	1973	1972	1969	1950
	U.	0	0	0	0	0	0	35 (1.756%)

Table 8 – part 1 of 2: Confusion matrix of evaluation of the passive sentence classifier (see Table 7).

P.A. – predicted active
A.A – actual active
P.P. – predicted passive
A.P.- actual passive
U. – Unclassified Instances
Note: Voted combination generates unclassified instances when the product of probabilities underflow.

Feature Set	Class	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
Euronews	P. P	36	89	66	84	67	50	29
	A. P.	53	0	23	5	22	39	5
	P. A.	55	1	12	9	20	47	4
	A. A.	1834	1888	1877	1880	1869	1842	1811
	U.	0	0	0	0	0	0	90 (4.550%)
Passive Sentences	P. P.	0	0	0	0	0	0	0
	A. P.	1098	1098	1098	1098	1098	1098	1098
	P. A.	1	1	1	1	1	1	1
	A. A.	0	0	0	0	0	0	0
	U.	0	0	0	0	0	0	0 (0.000%)

Table 8 – part 2 of 2: Confusion matrix of evaluation of the passive sentence classifier (see Table 7).

Feature Set	Class	Measure	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
All	Active	P	0.930	0.957	0.964	0.955	0.978	0.917	0.990
		R	0.953	0.968	0.967	0.970	0.965	0.954	0.986
		F1	0.942	0.962	0.965	0.962	0.971	0.935	0.988
	Passive	P	0.953	0.919	0.918	0.922	0.915	0.981	0.962
		R	0.963	0.891	0.909	0.887	0.946	0.965	0.974
		F1	0969	0.905	0.914	0.904	0.931	0.973	0.968

Table 9: Evaluation of the passive sentence classifier using 10-fold cross-validation and all features.
Note: We used a double weight to the passive dataset to have a comparable to have about the same size of the other datasets and increased less 1% F1 measures.

Feature Set	Class	S.M.O.	Bayesian Logistic Regression	Simple Logistic	Voted Perceptron	Simple Cart	C4.5	Voted
All	P. P.	166	252	211	263	125	106	53
	A. P.	5598	2069	2110	2058	2196	2215	1975
	P. A.	108	183	188	173	203	201	79
	A. A.	2213	5581	5576	5591	5561	5563	5445
	U.	0	0	0	0	0	0	533 (6.596%)

Table 10 – Confusion matrix of evaluation of the passive sentence classifier (see Table 9).

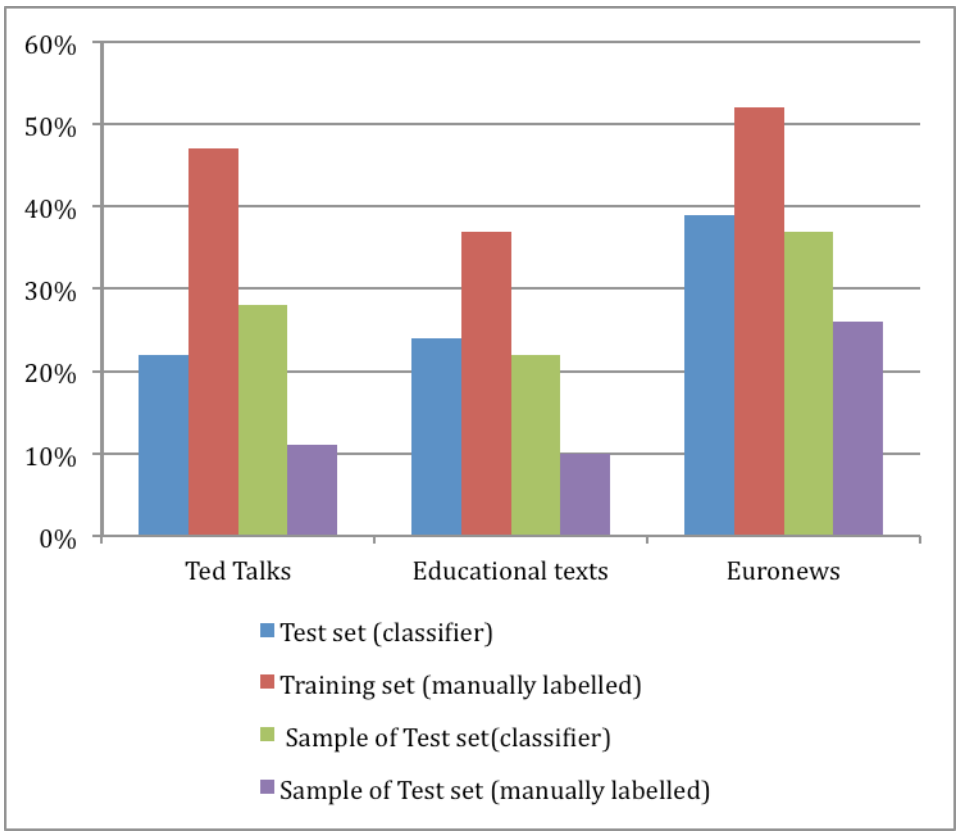

Fig. 2: Percentage of recommended sentence splitting for several text types.

4. DISCUSSION

We have seen that including the cue phrase position ratio features lead to a significant impact leading to an improvement, on average, between 0,1% and 2% using all training corpus (table 5) and between 3% and 5% using each training sets individually. We have also tried including more anaphoric cue words, such as possessive pronouns (mine, his, her, its, ours, yours, their), but gains resulting from their inclusion were not observed. The precision, recall, and F1 values were gather directly from Weka's output. Previous work gravitated around C4.5 and SMO classifiers; hence we started by testing both classifiers and processing testing and including other classifiers. Given the performance gap for several F1 measures of split class classifiers remaining below 10%, we decided to explore its combination by using a Meta classifier. The voted based classifier, where the product of probabilities of the classifiers was selected as the voted combination rule, yielded the best results. The results in table 4 and 5 confirmed our expectations. While table 4 provides the performance values of the classifiers on each dataset, table 5 confers an overview considering several combinations of groups of features. Upon inspection, SMO classifiers tend to behave slightly better than the other classifiers (excluding voted based classifier) when we include all training sets. We devised an explanation for this fact based on the amount of training data. Considering each training set individually, it is likely to be insufficient to be found the best splitting boundary by the SMO algorithm. The slightly better performance observed using solemnly the Ted Talks is also an important aspect that should be taken into account to decide which is the best type of text to proceed with text simplification work.

After analyzing several classifiers described, we observed that Voted Perceptron had a tendency to favor recall in detriment of the precision as its evaluation in the Euronews corpus reflects. Accordingly, we chose 20 epochs (iterations) for Voted Perceptron because we observed small improvements in terms of precision and recall above this number. While improvements above 20 epochs were residual, training time increased asymptotically as O(n), e.g. 1 epochs - took less than 10 minutes, 20 epochs near 3 hours, 200 epochs near 3 days, - the running times were estimated using a Intel Q9550 @ 2.83GHz, 3.3 GB memory ram - running Weka 3.6.2 under java 1.6.0 20).

Currently, the best performing sentence splitting classifier does not include a passive sentence classifier. Such fact was somehow expected. The typed dependency features already capture most of the information about passive sentences and we are accumulating errors from the passive classifier (F1 = 97,7%) and typed dependency features (F1 = 84.2%). In addition, the number of passive sentences is small in the training set which helps justifying the small

decrease in the classifier performance.

In our experiments, the passive classifier archived 97% in passive class F1 measure. Unfortunately, we cannot directly compare with an ordinary passive classifier (93% <= F1 <= 95%) [33]. Firstly, we had not access to the same training data. In addition, we decided to join reduced passives (typically recognized by parser as active voice) and mixed voice (both passive and active voice in the same sentence).

Further analysis justified the inclusion of the Passive sentence dataset (table 7). The low passive class F1 values (≤0.538) justify it.

The research goal is not only create the best state of art classifiers, but also go a step further and apply them to several types of text. Thus, next we will analyze the percentage of recommended splits. The percentage of recommended splits found in the training set are substantially higher than what is found in the sample of the sample of test set, despite the fact that both training and test sets share the same source. Thus, it might explain the discrepancy found between the manually labeled values and the classifier output for the sample.

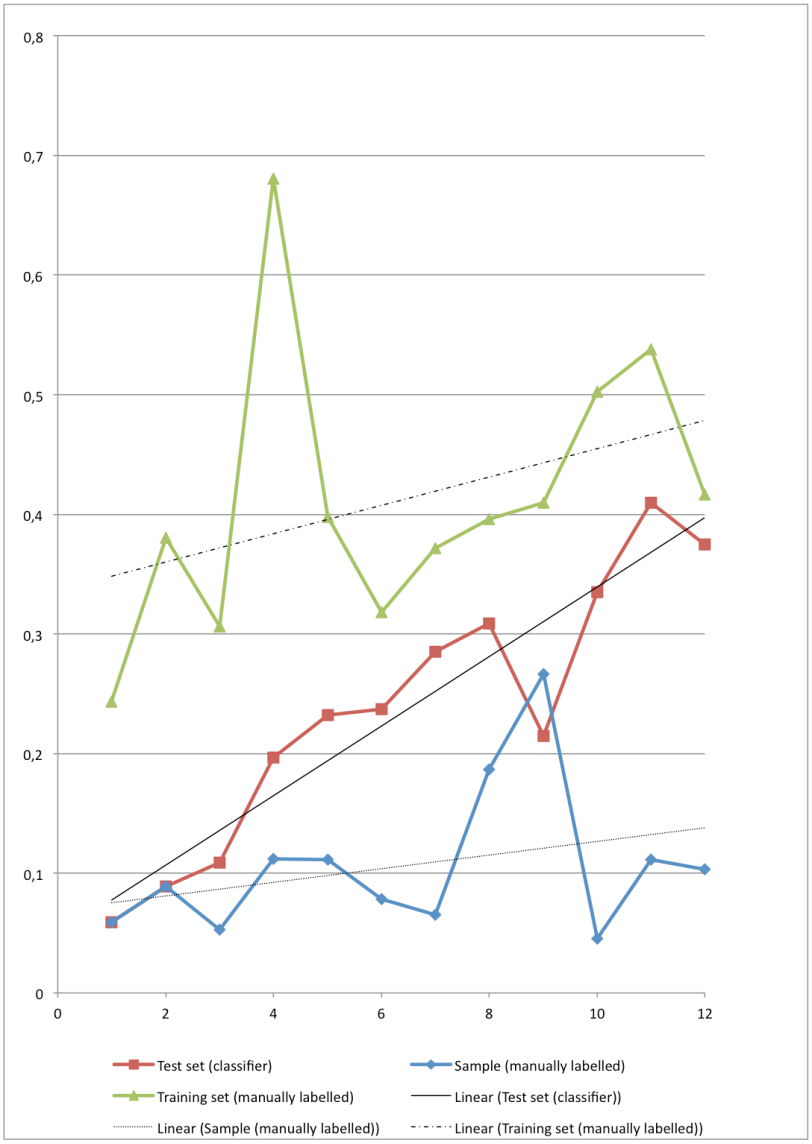

Fig. 3: Distribution of recommended Splits Per Scholastic Grade Level obtaining using the best voted based classifier.

The percentage of recommended sentence splits for Euronews texts in the test set is roughly the double of other text types as presented in Figure 2, because most of the news are summaries. Hence, the sentences tend to be longer and more complex. Unfortunately, it is not so preeminent in the other sets.

Despite the noise observed in Figure 3, we were able to draw 3 strict upward linear trendlines, correlating the recommended sentence splitting percentage and the readability levels. It is quite possible that by considering different text sizes we might be able to make this relation for types of text other than educational texts. On the other hand there is still further work to reduce that gap between trendlines. We would also like to see how the percentage of recommended splits for Ted Talks and higher readability level texts are related. In our experiments,

they had approximately the same percentage of splits found in the 8th grade texts. These results aroused our interest in using subtitles, such as Ted talks subtitles available in several languages, as source of text for automatic text simplification, given the existence of written texts by the same authors.

5. CONCLUSION

We have shown that by combining several classifiers and novel features, e.g. cue phrases position ratio, typed dependency features, we advanced the state of art of sentence splitting classifiers by comparison with previous work done for English (Medero and Ostendorf [13]) and Brazilian Portuguese (Gasperin et al. [12]). The passive classifier built during the process did not improve the spitting sentence classifier. Also, we noted that different kinds of texts have different percentage of recommended sentence splits, with summaries (Euronews) leading the list. We see an interesting relation between the readability level and the percentage of recommended sentence splits. However, further work analyzing the text length should be conducted.

Furthermore, we strove to create modular and language independent classifiers to allow not only its expansion, but also its easy portability to other languages beyond English.

This work brings up the possibility of creating text simplification systems targeting talks or movies subtitles; with a reduction of the cognitive load that makes the multimedia experience [31] very interesting for language technology-based education systems.

Future work should explore automatic processes to effectively split the sentences identified by the classifier in order to create a text simplification system.

6. ACKNOWLEDGMENTS

We would like to thank Alon Lavie, Isabel Trancoso, Maxine Eskenazi, and Nuno Mamede for fruitful discussions; Tiago Luís, João Miranda for their help in retrieving the Euronews and Ted talks corpora; and Kevin Dela Rosa for his help during evaluation of the gold standards.

ICTI CMU-Portugal provided support for this research. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors, and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] Luís Marujo, José Lopes, Nuno Mamede, Isabel Trancoso, Juan Pino, Maxine Eskenazi, Jorge Baptista, and Céu Viana, "Porting REAP to European Portuguese," in SLaTE 2009 - Speech and Language Technology in Education, Brighton, UK, 2009, Elsevier.

[2] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi, "Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension," Proceedings of the Ninth International Conference on Spoken Language Processing.

[3] R. Chandrasekar, Christine Doran, and B. Srinivas, "Automatic induction of rules for text simplification," In Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING 1996), pp. 1041–1044, 1996.

[4] Jia Xu, Richard Zens, and Hermann Ney, "Sentence Segmentation Using IBM Word Alignment Model 1," In Proceedings of EAMT 2005 (10th Annual Conference of the European Association for Machine Translation), pp. 280–287, 2005.

[5] Milton E. Goldman, "Using Captioned TV for Teaching Reading. FASTBACK 359.," 1993.

[6] Thomas J. Garza, "Evaluating the use of captioned video materials in advanced foreign language learning," Foreign Language Annals, vol. 24, pp. 239–258, 2008.

[7] Rebecca Oxford, "Learning a Language by Satellite Television: What Influences Student Achievement?.," System, vol. 21, no. 1, pp. 31–48, 1993.

[8] Hugo Meinedo, Márcio Viveiros, and Joa̋o Neto, "Evaluation of a live broadcast news subtitling system for Portuguese," in Interspeech 2008. 2008, ISCA.

[9] Gilles Boulianne Patrick Cardinal Claude Chapdelaine Michel Comeau Frederic Osterrath Pierre Ouellet Julie Brousseau, Jean-Francois Beaumont, "Automated Closed-Captioning of Live TV Broadcast News in French," in Eighth European Conference on Speech Communication and Technology. ISCA, 2003.

[10] Advaith Siddharthan, Syntactic simplification and text cohesion, Ph.D. thesis, University of Cambridge, 2004.

[11] Sarah E. Petersen and Mari Ostendorf, "Text simplification for language learners: a corpus analysis," in SLaTE 2007 - Speech and Language Technology in Education. 2007, pp. 69–72, Elsevier.

[12] Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Sandra M. Aluisio, "Learning when to simplify sentences for natural text simplification," Proceedings of ENIA, pp. 809–818, 2009.

[13] Julie Medero and Mari Ostendorf, "Analysis of vocabulary difficulty using Wiktionary," in SLaTE 2009 - Speech and Language Technology in Education, Brighton, UK, 2009, Elsevier.

[14] Kevin Lenzo,"The cmu pronouncing dictionary," 2010, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[15] Wikipedia, "Lists of common misspellings/for machines," 2010, http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines.

[16] "Apache lucene," 2010, http://lucene.apache.org/java/docs/index.html.

[17] Richard Landis and Gary Koch, "The measurement of observer agreement for categorical data," vol. 33, no. 1, pp.159–174, 1977.

[18] Philip R. "Pib" Burns, "Morphadorner," 2010, http://morphadorner.northwestern.edu/morphadorner/.

[19] Mitchell Marcus, Beatrice Santorini, and Mary Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," Association For Computational Linguistics, Jan 1994.

[20] Dan Klein and Christopher Manning, "Accurate Unlexicalized Parsing," Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423–430, 2003.

[21] Kristina Toutanova, Christopher Manning, and Yoram Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic De- pendency Network," Proceedings of HLT-NAACL 2003, pp. 252–259, 2003.

[22] Ian Witten and Eibe Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2005, 2nd edition.

[23] John C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Research Technical Report MSR-TR-98-14, 1998.

[24] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," IEEE Transactions Neural Computation, vol. 11, no. 5, pp. 1188–1194, 2000.

[25] Alexander Genkin, David D. Lewis, and David Madigan, "Large-scale Bayesian logistic regression for text categorization," Technical report, 2004.

[26] Niels Landwehr, Mark Hall, and Eibe Frank, "Logistic model trees," Machine Learning, vol. 59, no. 1, pp. 161–205,2005.

[27] Yoav Freund and Robert E. Schapire., "Large margin classification using the perceptron algorithm," Machine learning, vol. 37, no. 3, pp. 277–296, 1999.

[28] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone, "Classification and Regression Trees," 1984.

[29] Ross Quinlan, "C4.5: Programs for machine learning," Morgan Kaufmann Publishers, 1993.

[30] Josef kittler, Mohamad Hatef, Robert P.W. Duin, and Jiri Matas, "On Combining classifiers," IEEE Transactions of Patterns Analysis and Machine Learning, vol. 20, no. 3, pp. 226–239, 1993.

[31] John Sweller, "Implications of cognitive load theory for multimedia learning," The Cambridge handbook of multimedia learning, pp. 19–30, 2005.

[32] Sean Igo, "Identifying Reduced Passive Voice Constructions in Shallow Parsing Environments," M.S. thesis, The University of Utah, 2007.

[33] Sean Igo and Ellen Rilloff, "Learning to Identify Reduced Passive Verb Phrases with a Shallow Parser", In Proceedings of AAAI, pp. 1458–1461, 2008

[34] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006*.

[35]Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation*.