Strong Baselines for Complex Word Identification across Multiple Languages

Pierre Finnimore[1], **Elisabeth Fritzsch**[1], **Daniel King**[1], **Alison Sneyd**[1], **Aneeq Ur Rehman**[1], **Fernando Alva-Manchego**[1] and **Andreas Vlachos**[2]
[1]Department of Computer Science, University of Sheffield
[2]Department of Computer Science and Technology, University of Cambridge
{pmfinnimore,fritzsch.elisabeth,danielking1903}@gmail.com,
a.sneyd@shef.ac.uk,a.neeq8394@gmail.com,f.alva@shef.ac.uk,
andreas.vlachos@cst.cam.ac.uk

Abstract

Complex Word Identification (CWI) is the task of identifying which words or phrases in a sentence are difficult to understand by a target audience. The latest CWI Shared Task released data for two settings: monolingual (i.e. train and test in the same language) and cross-lingual (i.e. test in a language not seen during training). The best monolingual models relied on language-dependent features, which do not generalise in the cross-lingual setting, while the best cross-lingual model used neural networks with multi-task learning. In this paper, we present monolingual and cross-lingual CWI models that perform as well as (or better than) most models submitted to the latest CWI Shared Task. We show that carefully selected features and simple learning models can achieve state-of-the-art performance, and result in strong baselines for future development in this area. Finally, we discuss how inconsistencies in the annotation of the data can explain some of the results obtained.

1 Introduction

Complex Word Identification (CWI) consists of deciding which words (or phrases) in a text could be difficult to understand by a specific type of reader. In this work, we follow the CWI Shared Tasks (Paetzold and Specia, 2016; Yimam et al., 2018) and assume that a target word or multi-word expression (MWE[1]) in a sentence is given, and our goal is to determine if it is complex or not (an example is shown in Table 1). Under this setting, CWI is normally treated using supervised learning and feature engineering to build monolingual models (Paetzold and Specia, 2016; Yimam et al., 2018). Unfortunately, this approach is infeasible for languages with scarce resources of annotated data. In this paper, we are interested in both monolingual and cross-lingual CWI; in the latter, we build models to make predictions for languages not seen during training.

Sentence	Target word/MWE	Complex?
Both China and the Philippines flexed their muscles on Wednesday.	flexed	Yes
	flexed their muscles	Yes
	muscles	No

Table 1: An annotated sentence in the English dataset of the Second CWI Shared Task.

While monolingual CWI has been studied extensively (see a survey in Paetzold and Specia (2017)), the cross-lingual setup of the task was introduced only recently by Yimam et al. (2017b), who collected human annotations from native and non-native speakers of Spanish and German, and integrated them with similar data previously produced for three English domains (Yimam et al., 2017a): News, WikiNews and Wikipedia.

For the Second CWI Shared Task (Yimam et al., 2018), participants built monolingual models using the datasets previously described, and also tested their cross-lingual capabilities on newly collected French data. In the monolingual track, the best systems for English (Gooding and Kochmar, 2018) differed significantly in terms of feature set size and the model's complexity, to the best systems for German and Spanish (Kajiwara and Komachi, 2018). The latter used Random Forests with eight features, whilst the former used AdaBoost with 5000 estimators or ensemble voting combining AdaBoost and Random Forest classifiers, with about 20 features.

In the cross-lingual track, only two teams achieved better scores than the baseline: Kajiwara and Komachi (2018) who used length and frequency based features with Random Forests, and

[1]We consider n-grams with $n \geq 2$ as MWEs, while Yimam et al. (2018) used $n \geq 3$.

Proceedings of NAACL-HLT 2019, pages 970–977
Minneapolis, Minnesota, June 2 - June 7, 2019.

Bingel and Bjerva (2018) who implemented an ensemble of Random Forests and feed-forward neural networks in a multi-task learning architecture.

Our approach to CWI differs from previous work in that we begin by building competitive monolingual models, but using the same set of features and learning algorithm across languages. This reduces the possibility of getting high scores due to modelling annotation artifacts present in the dataset of one language. Our monolingual models achieve better scores for Spanish and German than the best systems in the Second CWI Shared Task. After that, we focus on language-independent features, and keep those that achieve good performance in cross-lingual experiments across all possible combinations of languages. This results in a small set of five language-independent features, which achieve a score as high as the top models in the French test set. Finally, we analyse the annotation of the datasets and find some inconsistencies that could explain some of our results.

Code for all our models can be found at: `https://github.com/sheffieldnlp/cwi`

2 Problem Formulation

We tackle the binary classification task in the Second CWI Shared Task (Yimam et al., 2018), in which a model decides if a target word/MWE in a sentence is complex or not. Following common practice, we extract features from the target word/MWE and its context, and then use a supervised learning algorithm to train a classifier. For training and testing our models, we use the annotated datasets provided for the Second CWI Shared Task (see Table 2 for some statistics).

Dataset	Train	Dev	Test
English (EN) - News	14,002	1,764	2,095
English (EN) - WikiNews	7,746	870	1,287
English (EN) - Wikipedia	5,551	694	870
Spanish (ES)	13,750	1,622	2,232
German (DE)	6,151	795	959
French (FR)	N/A	N/A	2,251

Table 2: Number of annotated samples in each dataset for each language.

3 Monolingual Models

3.1 Features Description

Our feature set consists of 25 features that can be extracted for all languages considered (English, German, Spanish and French). They can be divided into three broad categories: features based on the target word/MWE, sub-word level features, and sentence-level features to capture information from the target's context. As we intended that our features be applicable across languages, we drew on features found to be useful in previous work on CWI (Yimam et al., 2017b, 2018). We made use of the python libraries spaCy[2] (Honnibal and Montani, 2017) and NLTK[3] (Loper and Bird, 2002). Details on the resources used for extracting each feature can be found in Appendix A.

At the **target word/MWE level**, we experimented with features such as Named Entity (NE) type, part-of-speech, hypernym counts, number of tokens in the target, language-normalised number of characters in each word, and simple unigram probabilities. These features are linguistically motivated. The perceived complexity of a MWE may be higher than that of a single word, as each component word can be complex, or simple component words can be synthesised into a complex whole. Similarly, infrequent words are less familiar, so we would expect low-probability target words to be found more complex. Along these lines, proper nouns could be more complex, as there is a vast number of NEs, and the chance that a person has encountered any one of them is low. We would expect this trend to reverse for the NE type of organisations, in combination with the Enlgish-News dataset, as organisations mentioned in news articles are frequently global, and so the chance that a person has encountered a proper noun *that is an organisation* is often higher than for proper nouns in general. In total, 14 features were used at the target word/MWE level.

Our **sub-word level** features include prefixes, suffixes, the number of syllables, and the number of complex punctuation marks (i.e. punctuation within the target word/MWE, such as hyphens, that could denote added complexity). We would expect certain affixes to be useful features, as language users use sub-word particles like these to identify unknown words: by breaking up a word like "granted" into "grant-" and "-ed", readers can fall back on their knowledge of these component pieces to clarify the whole. A total of 9 sub-word features were used in the monolingual models.

Finally, **sentence level** features with linguistic

[2] `https://spacy.io/`
[3] `https://www.nltk.org/`

motivations were also considered. Long sentences could be harder to understand, which makes it more difficult to figure out the meaning of unknown words contained within them. Also, long sentences are more likely to include more unknown words or ambiguous references. Therefore, we considered sentence length (i.e., number of tokens in the sentence) as a feature. In addition, we extracted N-grams (unigrams, bigrams and trigrams) from the whole sentence, since certain sentence constructions can help a reader understand the target word/MWE. For example, "A of the B" suggests a relation between A and B. We used 2 sentence-level features in total.

3.2 Experiments and Results

Following Yimam et al. (2018), we used Macro-F1 score to evaluate performance and for comparison with previous work on the datasets. We used Logistic Regression for all our experiments, as it allowed for easy exploration of feature combinations, and in initial experiments we found that it performed better than Random Forests. We evaluated both using the full feature set described before, as well as a two-feature baseline using the number of tokens of the target and its language-normalised number of characters. Results of our monolingual experiments are shown in Table 3.

Dataset	Dev		Test		
	BL	MA	BL	MA	SotA
EN - News	83.6	85.5	69.7	86.0	**87.4**
EN - WikiNews	80.4	82.8	65.8	81.6	**84.0**
EN - Wikipedia	74.2	76.6	70.1	76.1	**81.2**
ES	78.0	77.1	69.6	**77.6**	77.0
DE	79.5	74.6	72.4	74.8	**75.5**
Mean	79.1	79.3	69.5	79.2	N/A

Table 3: Macro-F1 for the baseline (BL), our monolingual approach (MA), and the state of the art (SotA) on the Dev and Test splits of each dataset.

In the test set, our baseline results (BL in Table 3) are strong, especially in German. Our full 25-features model improves on the baseline in all cases, with the biggest increase of over 16 percentage points seen for the EN-News dataset. Our system beats the best performing system from the Shared Task in Spanish (77.0) and German (74.5), both obtained by Kajiwara and Komachi (2018). However, the state of the art for German remains the Shared Task baseline (75.5) (Yimam et al., 2018). The best results for all three English datasets were obtained by Gooding and Kochmar (2018); ours is within two percentage points of their News dataset score. Furthermore, the mean score for our system (79.2) is close to the mean of the best performing models (81.0), which are different systems, while using simpler features and learning algorithm. The best-performing model in English, for example, used Adaboost with 5000 estimators (Gooding and Kochmar, 2018).

4 Cross-lingual Models

4.1 Features Description

Linguistically, the cross-lingual approach can be motivated by the relation between certain languages (such as French and Spanish both being Romance languages). In addition, there may be features identifying complex words that are shared even across language families.

To be able to test a model on a language that was unseen during training, the features the model works with must be cross-lingual (or language-independent) themselves. For example, the words themselves are unlikely to transfer across languages (apart from those that happen to be spelled identically), but the popularity of the words would transfer. This rules out some of the features we used for the monolingual approach (see Sec. 3.1), as they were language-dependent. One such feature is N-grams for the target word/MWE, which depend on the language, and so will only occur with extreme sparsity outside of their source language. For example, if applying a system trained on English to unseen French, the English phrases "à la mode" or "film noir" might reoccur in the French, since they originate from that language, but these are rare exceptions. What is more, a French loan-phrase may have different complexity characteristics to the same N-grams occurring in their native language. Therefore, we did not use these features in the cross-lingual system.

4.2 Experiments and Results

To find out which features were best suited for the cross-lingual approach, we performed an iterative ablation analysis (see Appendix B for details). Using this process, we arrived at our final cross-lingual feature set: number of syllables in the target, number of tokens in the target, number of complex punctuation marks (such as hyphens), sentence length, and unigram probabilities.

Furthermore, we analyse the effect of different language combinations on the performance of the cross-lingual model in order to investigate how the relationship between the languages trained and tested on would influence model performance. Recall that we only have training data for English, Spanish and German, but not French. We train models using all possible combinations (each language independently, each pairing, and all three) and evaluate on each of the four languages that have test data (i.e. the former three and French), excluding training combinations that include the test language. Results are shown in Table 4.

EN	ES	DE	Eval	Source	Test	Dev
	✓	✓	EN	WikiNews	61.8	63.7
		✓	EN	WikiNews	62.3	63.6
	✓		EN	WikiNews	61.6	63.8
	✓	✓	EN	Wikipedia	62.8	64.4
		✓	EN	Wikipedia	62.6	64.4
	✓		EN	Wikipedia	63.1	65.2
	✓	✓	EN	News	67.1	65.6
		✓	EN	News	67.0	65.6
	✓		EN	News	**67.2**	**65.9**
✓		✓	ES	N/A	70.8	71.3
		✓	ES	N/A	**72.6**	**74.1**
✓			ES	N/A	69.1	70.0
✓	✓		DE	N/A	**73.4**	**78.3**
	✓		DE	N/A	72.6	77.4
✓			DE	N/A	73.0	76.0
✓	✓	✓	FR	N/A	73.1	N/A
	✓	✓	FR	N/A	75.7	N/A
✓	✓		FR	N/A	73.4	N/A
✓		✓	FR	N/A	70.5	N/A
		✓	FR	N/A	**75.8**	N/A
	✓		FR	N/A	73.4	N/A
✓			FR	N/A	69.2	N/A

Table 4: Comparison of Test and Dev results for all permutations of training languages.

When testing on French, we achieved the highest performance by training on German only (75.8), followed closely by training on a combination of German and Spanish (75.7) and only Spanish (75.5). The worst performance was achieved by training only on English (69.2), and the performance also noticeably decreased for all training combinations that included English.

When testing on German, language choice had a weaker effect. The highest score came from combining English and Spanish (73.4), but using only one of those languages gave comparable results (72.6 for Spanish, 73.0 for English).

For Spanish, the best results were achieved when training only on German (72.6). Adding English to the training languages decreased the performance (70.8), which was even lower when training only on English (69.1).

	Spanish	German	French
Monolingual SotA	**77.0**	**75.5**	N/A
Cross-lingual SotA	N/A	N/A	**76.0**
Our cross-lingual	72.6	73.4	75.8

Table 5: Comparison between the monolingual and cross-lingual state of the art (SotA), and our cross-lingual system.

It is noteworthy that adding English to the training languages noticeably decreases performance for both Spanish and French, but not for German. One possible reason for Spanish and French not benefiting from English when German does is that both English and German are Germanic languages, whereas Spanish and French are Romance languages. Another possible explanation for the decrease of performance caused by training with English is that there are inconsistencies in the way MWEs in the datasets were labelled across languages, which we explore in Sec. 5.

We finally compare our cross-lingual models against the state of the art: the best monolingual system for Spanish and German, and the best cross-lingual system for French, where no monolingual systems exist. As Table 5 shows, our cross-lingual models come close to the best monolingual models for Spanish and especially for German. This is remarkable given how simple our model and features are, and that the approaches we compare against train complex models for each language. Furthermore, this points towards the possibility of extending CWI to more languages which lack training data.

Finally, Table 6 compares the coefficients for models trained on Romance and Germanic languages. Notably, use of complex punctuation (such as the hyphenation in “laser-activated” or “drug-related”) and the number of tokens are inversely correlated w.r.t. the word or MWE being complex. More words in the target was correlated with complexity for English and German, and inversely correlated for Spanish.

5 Dataset Analysis

While examining our models’ incorrect predictions, we observed inconsistencies in labelling in the datasets between target MWEs and their subwords/sub-expressions (SWs).

Feature	Train	Coefficient
number of complex punctuation marks	EN	-0.693
	DE	-0.559
	ES	1.111
number of tokens	EN	-2.200
	DE	-0.534
	ES	1.420

Table 6: Coefficients for cross-lingual models trained on Germanic and Romance languages.

The First CWI Shared Task (Paetzold and Specia, 2016) used the annotations of a group (i.e. ten annotators on the training data) to predict the annotation of an individual (i.e. one annotator on the test data). The resulting inconsistencies in labelling may have contributed to the low F-scores of systems in the task (Zampieri et al., 2017). Although the Second CWI Shared Task improved on the first by having multiple annotators for all splits of the data, it contains some labelling inconsistencies arising from annotators now being able to label phrases, and words within them, separately.

More concretely, we found that across all datasets, 72% of MWEs contain at least one SW with the opposite label (see Table 7). While this makes sense in some cases, every SW in 25% of MWE instances has the opposite label. For example, "numerous falsifications and ballot stuffing" is not annotated as complex, despite its SWs "numerous", "numerous falsifications", "falsifications", "ballot", "ballot stuffing" and "stuffing" all being complex. Conversely, "crise des marchés du crédit" is complex, despite "crise", "marchés" and "crédit" being labelled non-complex. It is difficult to see how classifiers that extract features for MWEs from their individual SWs could predict the labels of both correctly.

Furthermore, every target MWE in the Spanish, German and French datasets is labelled complex. This may bias a classifier trained on the Spanish or German datasets towards learning MWEs and long individual words (if length is a feature) are complex. In particular, this observation may help explain why adding English as a training language decreased the performance of our cross-lingual system when testing on French and Spanish (where all MWEs are complex). An analysis in Bingel and Bjerva (2018) further found that their cross-lingual French model was effective at predicting long complex words/MWEs but had difficulty predicting long non-complex words.

	C	NC	$\geq$ 1 Irreg.	All Irreg.
English	3,750	982	3,315	950
Spanish	2,309	0	1,747	760
German	502	0	374	178
French	242	0	192	82
Total	6,803	982	5,628	1,970

Table 7: MWE annotation analysis: numbers of MWEs labelled complex (C) and non-complex (NC), numbers with at least one SW ($\geq$ 1 Irreg) and all SWs (All Irreg.) having the opposite label.

It is also worth noting that considering a word or MWE as complex is subjective and may differ from person to person, even within the same target audience. Bingel et al. (2018) investigated predicting complex words based on the gaze patterns of children with reading difficulties. They found a high degree of specificity in misreadings between children, that is, which words they found complex when reading aloud. This variety of complexity judgements even within one target group points towards the high degree of subjectivity in the task, which may also partly explain the inconsistencies in the dataset.

6 Conclusion and Future Work

The monolingual and cross-lingual models presented achieve comparable results against more complex, language-specific state-of-the-art models, and thus can serve as strong baselines for future research in CWI. In addition, our analysis of the dataset could help in the design of better guidelines when crowdsourcing annotations for the task. Dataset creators may wish to only allow single words to be chosen as complex to avoid labelling inconsistencies. In case MWEs are being permitted, we suggest instructing annotators to chose the smallest part of a phrase they find complex (French annotators for the Second CWI Shared Task sometimes grouped individual complex words into a complex MWE (Yimam et al., 2018)).

Acknowledgements

This work was initiated in a class project for the NLP module at the University of Sheffield. The authors would like to acknowledge the contributions of Thomas Dakin, Sanjana Khot and Harry Wells who contributed their project code to this work. Andreas Vlachos is supported by the EPSRC grant eNeMILP (EP/R021643/1).

References

Joachim Bingel, Maria Barrett, and Sigrid Klerke. 2018. Predicting misreadings from gaze in children with reading difficulties. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 24–34, New Orleans, Louisiana. Association for Computational Linguistics.

Joachim Bingel and Johannes Bjerva. 2018. Cross-lingual complex word identification with multitask learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 166–174. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2018. Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569. Association for Computational Linguistics.

Gustavo H. Paetzold and Lucia Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78. Association for Computational Linguistics.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017a. Cwig3g2 - complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407. Asian Federation of Natural Language Processing.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017b. Multilingual and cross-lingual complex word identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 813–822. INCOMA Ltd.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex word identification: Challenges in data annotation and system performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 59–63. Asian Federation of Natural Language Processing.

A Detailed Feature Set

Level	Name	Description	Resource
Target word/MWE	NER_tag_counts	Counts of each Named Entity tag in target	spaCy
	pos_tag_counts	Counts of each part-of-speech tag in target	spaCy
	hypernym_count	Number of hypernyms	WordNet (NLTK)
	len_tokens	Absolute length in tokens	N/A
	len_tokens_norm	Normalised length in tokens	N/A
	len_chars_norm	Normalised length in characters	N/A
	unigram_prob	Log of the product of unigram probabilities	EN: Brown Corpus (NLTK) ES: CESS-ESP (NLTK) DE: TIGER Corpus[4] FR: Europal[5]
	bag_of_shapes	Bag of morphological shapes	spaCy
	rare_word_count	Count of rare words in target	EN: subset of Google's Trillion Word Corpus[6] DE: list of the most common 3,000 words[7] ES: word frequency list by M. Buchmeier[8]
	rare_trigram_count	Count of rare trigrams in target	Same as rare_word_count
	is_stop	Frequency of stopwords in target	NLTK, Ranks NL[9]
	is_nounphrase	If target is a noun phrase	spaCy
	avg_chars_per_word	Avg. word length (in characters) of the target	N/A
	iob_tags	Count of BIO tags in target	spaCy
Sub-word	lemma_feats	Bag of lemmas for target sentence	spaCy
	len_sylls	Length of target in syllables	Pyphen[10]
	num_complex_punct	Count of complex punctuation in target	N/A
	char_n_gram_feats	Character N-Grams, incl. prefixes and suffixes	N/A
	char_tri_sum	Sum of character trigrams' corpus frequencies	EN: Brown Corpus (NLTK) ES: CESS-ESP (NLTK) DE: TIGER Corpus
	char_tri_avg	Average of character trigrams' corpus frequencies	same as char_tri_sum
	consonant_freq	Count of consonants in target	N/A
	gr_or_lat	If target has Greek or Latin affixes	List of Greek and Latin roots in English[11]
	is_capitalised	If target's first letter is uppercased	N/A
Sentence	sent_length	Number of tokens in the sentence	N/A
	sent_n_gram_feats	Unigrams, bigrams and trigrams in the sentence	N/A

Table 8: Monolingual and Cross-lingual Feature Set Summary

[4] https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
[5] http://www.statmt.org/europarl/
[6] https://github.com/first20hours/google-10000-english
[7] http://germanvocab.com/
[8] https://en.wiktionary.org/wiki/User:Matthias_Buchmeier/Spanish_frequency_list-1-5000
[9] https://www.ranks.nl/stopwords
[10] https://pyphen.org/
[11] https://www.oakton.edu/user/3/gherrera/Greek%20and%20Latin%20Roots%20in%20English/greek_and_latin_roots.pdf

B Cross-lingual Features Ablation

Iteration	Current features	Features increasing performance	Features decreasing performance
1	len_tokens	num_complex_punct len_sylls sent_length	is_nounphrase len_tokens_norm consonant_freq is_capitalised bag_of_shapes pos_tag_count
2	len_tokens len_sylls num_complex_punct sent_length	unigram_prob	gr_or_lat
3	len_tokens len_sylls num_complex_punct sent_length unigram_prob		char_ngram_feats iob_tags lemma_feats NER_tag_counts

Table 9: Ablation analysis for the cross-lingual features