Learning Simple Wikipedia: A Cogitation in Ascertaining Abecedarian Language

Courtney Napoles and **Mark Dredze**
Center for Language and Speech Processing
Human Language Technology Center of Excellence
Johns Hopkins University
Baltimore, MD 21211
courtneyn@jhu.edu, mdredze@cs.jhu.edu

Abstract

Text simplification is the process of changing vocabulary and grammatical structure to create a more accessible version of the text while maintaining the underlying information and content. Automated tools for text simplification are a practical way to make large corpora of text accessible to a wider audience lacking high levels of fluency in the corpus language. In this work, we investigate the potential of Simple Wikipedia to assist automatic text simplification by building a statistical classification system that discriminates *simple* English from *ordinary* English. Most text simplification systems are based on hand-written rules (e.g., PEST (Carroll et al., 1999) and its module SYSTAR (Canning et al., 2000)), and therefore face limitations scaling and transferring across domains. The potential for using Simple Wikipedia for text simplification is significant; it contains nearly 60,000 articles with revision histories and aligned articles to ordinary English Wikipedia. Using articles from Simple Wikipedia and ordinary Wikipedia, we evaluated different classifiers and feature sets to identify the most discriminative features of simple English for use across domains. These findings help further understanding of what makes text simple and can be applied as a tool to help writers craft simple text.

1 Introduction

The availability of large collections of electronic texts is a boon to information seekers, however, advanced texts often require fluency in the language. Text simplification (TS) is an emerging area of text-to-text generation that focuses on increasing the readability of a given text. Potential applications can increase the accessibility of text, which has great value in education, public health, and safety, and can aid natural language processing tasks such as machine translation and text generation.

Corresponding to these applications, TS can be broken down into two rough categories depending on the target "reader." The first type of TS aims to increase human readability for people lacking high-level language skills, either because of age, education level, unfamiliarity with the language, or disability. Historically, generating this text has been done by hand, which is time consuming and expensive, especially when dealing with material that requires expertise, such as legal documents. Most current automatic TS systems rely on handwritten rules, e.g., PEST (Carroll et al., 1999), its SYSTAR module (Canning et al., 2000), and the method described by Siddharthan (2006). Systems using handwritten rules can be susceptible to changes in domains and need to be modified for each new domain or language. There has been some research into automatically learning the rules for simplifying text using aligned corpora (Daelemans et al., 2004; Yatskar et al., 2010), but these have yet to match the performance hand-crafted rule systems. An example of a manually simplified sentence can be found in table 1.

The second type of TS has the goal of increasing the machine readability of text to aid tasks such as information extraction, machine translation, generative summarization, and other text generation

tasks for selecting and evaluating the best candidate output text. In machine translation, the evaluation tool most commonly used for evaluating output, the BLEU score (Papineni et al., 2001), rates the "goodness" of output based on n-gram overlap with human-generated text. However this metric has been criticized for not accurately measuring the fluency of text and there is active research into other metrics (Callison-Burch et al., 2006; Ye et al., 2007). Previous studies suggest that text simplified for machine and human comprehension are categorically different (Chae and Nenkova, 2009). Our research considers text simplified for human readers, but the findings can be used to identify features that discriminate simple text for both applications.

The process of TS can be divided into three aspects: removing extraneous or superfluous text, substituting more complex lexical and syntactic forms, and inserting information to offer further clarification where needed (Aluísio et al., 2008). In this regard, TS is related to several different natural language processing tasks such as text summarization, compression, machine translation, and paraphrasing.

While none of these tasks alone directly provide a solution to text simplification, techniques can be drawn from each. Summarization techniques can be used to identify the crucial, most informative parts of a text and compression can be used to remove superfluous words and phrases. In fact, in the Wikipedia documents analyzed for this research, the average length of a "simple" document is only 21% the length of an "ordinary" English document (although this may be an unintentional byproduct of how articles were simplified, as discussed in section 6.1).

In this paper we study the properties of language that differentiate simple from ordinary text for human readers. Specifically, we use statistical learning techniques to identify the most discriminative features of simple English and "ordinary" English using articles from Simple Wikipedia and English Wikipedia. We use cognitively motivated features as well as statistical measurements of a document's lexical, syntactic, and surface features. Our study demonstrates the validity and potential benefits of using Simple Wikipedia as a resource for TS research.

Ordinary text
Every person has the right to a name, in which is included a first name and surname. ... The alias chosen for legal activities has the same protection as given to the name.
Same text in simple language
Every person has the right to have a name, and the law protects people's names. Also, the law protects a person's alias. ... The name is made up of a first name and a surname (name = first name + surname).

Table 1: A text in ordinary and simple language from Aluísio et al. (2008).

2 Wikipedia as a Corpus

Wikipedia is a unique resource for natural language processing tasks due to its sheer size, accessibility, language diversity, article structure, inter-document links, and inter-language document alignments. Denoyer and Gallinari (2006) introduced the Wikipedia XML Corpus, with 1.5 million documents in eight languages from Wikipedia, that stored the rich structural information of Wikipedia with XML. This corpus was designed specifically for XML retrieval but has uses in natural language processing, categorization, machine translation, entity ranking, etc. YAWN (Schenkel et al., 2007), a Wikipedia XML corpus with semantic tags, is another example of exploiting Wikipedia's structural information. Wikipedia provides XML site dumps every few weeks in all languages as well as static HTML dumps.

A diverse array of NLP research in the past few years has used Wikipedia, such as for word sense disambiguation (Mihalcea, 2007), classification (Gantner and Schmidt-Thieme, 2009), machine translation (Smith et al., 2010), coreference resolution (Versley et al., 2008; Yang and Su, 2007), sentence extraction for summarization (Biadsy et al., 2008), information retrieval (Muller and Gurevych, 2008), and semantic role labeling (Ponzetto and Strube, 2006), to name a few. However, except for very recent work by Yatskar et al. (2010), to our knowledge there has not been comparable research in using Wikipedia for text simplification.

What makes Wikipedia an excellent resource for

text simplification is the new Simple Wikipedia project[1], a collection of 58,000 English Wikipedia articles that have been rewritten in Simple English, which uses basic vocabulary and less complex grammar to make the content of Wikipedia accessible to students, children, adults with learning difficulties, and non-native English speakers. In addition to being a large corpus, these articles are linked to their ordinary Wikipedia counterparts, so for each article both a simple and an ordinary version are available. Furthermore, on inspection many articles in Simple Wikipedia appear to be copied and edited from the corresponding ordinary Wikipedia article. This information, together with revision history and flags signifying unsimplified text, can provide a scale of information on the text-simplification process previously unavailable. Example sentences from Simple Wikipedia and ordinary Wikipedia are shown in table 2.

We used articles from Simple Wikipedia and ordinary English Wikipedia to create a large corpus of simple and ordinary articles for our experiments. In order to experiment with models that work across domains, the corpus includes articles from nine of the primary categories identified in Simple Wikipedia: Everyday Life, Geography, History, Knowledge, Literature, Media, People, Religion, and Science. A total of 55,433 ordinary and 42,973 simple articles were extracted and processed from English Wikipedia and Simple Wikipedia, respectively. Each document contains at least two sentences. Additionally, the corpus contains only the main text body of each article and does not consider info boxes, tables, lists, external and cross-references, and other structural features. The experiments that follow randomly extract documents and sentences from this collection.

Before extracting features, we ran a series of natural language processing tools to preprocess the collection. First, all of the XML and "wiki markup" was removed. Each document was split into sentences using the Punkt sentence tokenizer (Kiss and Strunk, 2006) in NLTK (Bird and Loper, 2004). We then parsed each sentence using the PCFG parser of Huang and Harper (2009), a modified version of the Berkeley parser (Petrov et al., 2006; Petrov and Klein, 2007), for the tree structure and part-of-speech tags.

[1] http://simple.wikipedia.org/

Ordinary Wikipedia
Hawking was the Lucasian Professor of Mathematics at the University of Cambridge for thirty years, taking up the post in 1979 and retiring on 1 October 2009.
Simple Wikipedia
Hawking was a professor of mathematics at the University of Cambridge (a position that Isaac Newton once had). He retired on October 1st 2009.

Table 2: Comparable sentences from the ordinary Wikipedia and Simple Wikipedia entry for "Stephen Hawking."

Coarse Tag	**Penn Treebank Tags**
DET	DT, PDT
ADJ	JJ, JJR, JJS
N	NN, NNS, NP, NPS, PRP, FW
ADV	RB, RBR, RBS
V	VB, VBN, VBG, VBP, VBZ, MD
WH	WDT, WP, WP$, WRB

Table 3: A mapping of the Penn Treebank tags to a coarse tagset used to generate features.

3 Task Setup

To evaluate the feasibility of learning simple and ordinary texts, we sought to identify text properties that differentiated between these classes. Using the two document collections, we constructed a simple binary classification task: label a piece of text as either simple or ordinary. The text was labeled according to its source: simple or ordinary Wikipedia. From each piece of text, we extracted a set of features designed to capture differences between the texts, using cognitively motivated features based on a document's lexical, syntactic, and surface features. We first describe our features and then our experimental setup.

4 Features

We began by examining the guidelines for writing Simple Wikipedia pages.[2] These guidelines suggest that articles use only the 1000 most common and basic English words and contain simple grammar and short sentences. Articles should be short but can be longer if they need to explain vocabulary words necessary to understand the topic. Additionally, words should appear on lists of basic English words, such as the Voice of America Special English words list (Voice Of America, 2009) or the Ogden Basic English list (Ogden, 1930). Idioms should be avoided as well as compounds and the passive voice as opposed to a single simple verb.

To capture these properties in the text, we created four classes of features: lexical, part-of-speech, surface, and parse. Several of our features have previously been used for measuring text fluency (Aluísio et al., 2008; Chae and Nenkova, 2009; Feng et al., 2009; Petersen and Ostendorf, 2007).

Lexical. Previous work by Feng et al. (2009) suggests that the document vocabulary is a good predictor of document readability. Simple texts are more likely to use basic words more often as opposed to more complicated, domain-specific words used in ordinary texts. To capture these features we used a unigram bag-of-words representation. We note that lexical features are unlikely to be useful unless we have access to a large training corpus that allowed the estimation of the relative frequency of words (Chae and Nenkova, 2009). Additionally, we can expect lexical features to be very fragile for cross-domain experiments as they are especially susceptible to changes in domain vocabulary. Nevertheless, we include these features as a baseline in our experiments.

Parts of speech. A clear focus of the simple text guidelines is grammar and word type. One way of representing this information is by measuring the relative frequency of different types of parts of speech. We consider simple unigram part-of-speech tag information. We measured the normalized counts and relative frequency of part-of-speech tags and counts of bigram part-of-speech tags in each piece of text. Since Devlin and Unthank (2006) has shown that word order (subject verb object (SVO), object verb subject (OVS), etc.) is correlated with readability, we also included a reduced tagset to capture grammatical patterns (table 3). We also included normalized counts of these reduced tags in the model.

Feature	Simple	Ordinary
Tokens	158	4332
Types	100	1446
Sentences	10	172
Average sentence length	15.80	25.19
Type-token ratio	0.63	0.33
Percent simple words	0.31	0.08
Not BE850 type-token ratio	0.65	0.30
BE850 type-token ratio	0.59	0.67

Table 4: A comparison of the article "Stephen Hawking" from Simple and ordinary Wikipedia.

Surface features. While lexical items may be important, more general properties can be extracted from the lexical forms. We can also include features that correspond to surface information in the text. These features include document length, sentence length, word length, numbers of lexical types and tokens, and the ratio of types to tokens. All words are labeled as basic or not basic according to Ogden's Basic English 850 (BE850) list (Ogden, 1930).[3] In order to measure the lexical complexity of a document, we include features for the number of BE850 words, the ratio of BE850 words to total words, and the type-token ratio of BE850 and non-BE850 words. Investigating the frequency and productivity of words not in the BE850 list will hopefully improve the flexibility of our model to work across domains and not learn any particular jargon. We also hope that the relative frequency and productivity measures of simple and non-simple words will codify the lexical choices of a sentence while avoiding the aforementioned problems with including specific lexical items.

[2] http://simple.wikipedia.org/wiki/Wikipedia:Simple_English_Wikipedia

[3] Wikipedia advocates using words that appear on the BE850 list. Ogden also provides extended Basic English vocabulary lists, totaling 2000 Basic English words, but these words tend to be more specialized or domain specific. For the purposes of this study only words in BE850 were used.

Table 4 shows the difference in some surface statistics in an aligned document from Simple and ordinary Wikipedia. In this example, nearly one-third of the words in the simple document are from the BE850 while less than a tenth of the words in the ordinary document are. Additionally, the productivity of words, particularly non-BE850 words, is much higher in the ordinary document. There are also clear differences in the length of the documents, and on average documents from ordinary Wikipedia are more than four times longer than documents from Simple Wikipedia.

Syntactic parse. As previously mentioned, a number of Wikipedia's writing guidelines focus on general grammatical rules of sentence structure. Evidence of these rules may be captured in the syntactic parse of the sentences in the text. Chae and Nenkova (2009) studied text fluency in the context of machine translation and found strong correlations between parse tree structures and sentence fluency.

In order to represent the structural complexity of the text, we collected extracted features from the parse trees. Our features included the frequency and length of noun phrases, verb phrases, prepositional phrases, and relative clauses (including embedded structures). We also considered relative ratios, such as the ratio of noun to verb phrases, prepositional to noun phrases, and relative clauses to noun phrases. We used the length of the longest noun phrase as a signal of complexity, and we also sought features that measured how typical the sentences were of English text. We included some of the features from the parser reranking work of Charniak and Johnson (2005): the height of the parse tree and the number of right branches from the root of the tree to the furthest right leaf that is not punctuation.

5 Experiments

Using the feature sets described above, we evaluated a simple/ordinary text classifier in several settings on each category. First, we considered the task of document classification, where a classifier determines whether a full Wikipedia article was from ordinary English Wikipedia or Simple Wikipedia. For each category of articles, we measured accuracy on this binary classification task using 10-fold cross-validation. In the second setting, we considered the performance of a sentence-level classifier. The classifier labeled each sentence as either ordinary or simple and we report results using 10-fold cross-validation on a random split of the sentences. For both settings we also evaluated a single classifier trained on all categories.

Category	**Documents**	**Sentences**
Everyday Life	15,124	7,392
Geography	10,470	5,852
History	5,174	1,644
Literature	992	438
Media	502	429
People	4,326	1,562
Religion	1,863	1,581
Science	25,787	21,054
All	64,238	39,952

Table 5: The number of examples available in each category. To compare experiments in each category we used at most 2000 instances in each experiment.

Feature class		**Features**
Lexical		**522,153**
Part of speech		**2478**
	tags	45
	tag pairs	1972
	tags (reduced)	22
	tag pairs (reduced)	484
Parse		**11**
Surface		**9**

Table 6: The number of features in each feature class.

We next considered cross-category performance: how would a classifier trained to detect differences between simple and ordinary examples from one category do when tested on another category. In this experiment, we trained a single classifier on data from a single category and used the classifier to label examples from each of the other categories. We report the accuracy on each category in these transfer experiments.

For learning we require a binary classifier training algorithm. We evaluated several learning algorithms for classification and report results for each one: a) MIRA—a large margin online learning algorithm (Crammer et al., 2006). Online learning algorithms observe examples sequentially and update the current hypothesis after each observation; b)

Confidence Weighted (CW) learning—a probabilistic large margin online learning algorithm (Dredze et al., 2008); c) Maximum Entropy—a log-linear discriminative classifier (Berger et al., 1996); and d) Support Vector Machines (SVM)—a large margin discriminator (Joachims, 1998).

For each experiment, we used default settings of the parameters and 10 online iterations for the online methods (MIRA, CW). To create a fair comparison for each category, we limited the number of examples to a maximum of 2000.

6 Results

For the first task of document classification, we saw at least 90% mean accuracy with each of the classifiers. Using all features, SVM and Maximum Entropy performed almost perfectly. The online classifiers, CW and MIRA, displayed similar preference to the larger feature sets, lexical and part-of-speech counts. When using just lexical counts, both CW and MIRA were more accurate than the SVM and Maximum Entropy (reporting 92.95% and 86.55% versus 75.00% and 78.75%, respectively). For all classifiers, the models using the counts of part-of-speech tags did better than classifiers trained on the surface features and on the parse features. This is surprising, since we expected the surface features to be robust predictors of the document class, mainly because the average ordinary Wikipedia article in our corpus is about four times longer than the average Simple Wikipedia article. We also expected the syntactic features to be a strong predictor of the document class since more complicated parse trees correspond to more complex sentences.

For each classifier, we looked at its performance without its less predictive feature categories, and for CW the inclusion of the surface features decreased performance noticeably. The best CW classifiers used either part-of-speech and lexical features (95.95%) or just part-of-speech features (95.80%). The parse features, which by themselves only yielded 64.60% accuracy, when combined with part-of-speech and lexical features showed high accuracy as well (95.60%). MIRA also showed higher accuracy when surface features were not included (from 97.50% mean accuracy with all features to 97.75% with all but surface features).

The best SVM classifier used all four feature classes, but had nearly as good accuracy with just part-of-speech counts and surface features (99.85% mean accuracy) and with surface and parse features (also 99.85% accuracy). Maximum Entropy, on the other hand, improved slightly when the lexical and parse features were not included (from 99.45% mean accuracy with all feature classes to 99.55%).

We examined the weights learned by the classifiers to determine the features that were effective for learning. We selected the features with the highest absolute weight for a MIRA classifier trained on all categories. The most predictive features for document classification were the sentence length (shorter favors Simple), the length of the longest NP (longer favors ordinary), the number of sentences (more favors ordinary), the average number of prepositional phrases and noun phrases per sentence, the height of the parse tree, and the number of adjectives. The most predictive features for sentence classification were the ratio of different tree non-terminals (VP, S, NP, S-Bar) to the number of words in the sentence, the ratio of the total height of the productions in a tree to the height of the tree, and the extent to which the tree was right branching. These features are consistent with the rules described above for simple text.

Next we looked at a pairwise comparison of how the classifiers perform when trained on one category and tested on another. Surprisingly, the results were robust across categories, across classifiers. Using the best feature class as determined in the first task, the average drop in accuracy when trained on each domain was very low across all classifiers (the mean accuracy rate of each cross-category classification was at least 90%). Table 6 shows the mean change in accuracy from CW models trained and tested on the same category to the models trained and tested on different categories. When trained on the Everyday Life category, the model actually showed a mean increase in accuracy when predicting other categories.

In the final task, we trained binary classifiers to identify simple sentences in isolation. The mean accuracy was lower for this task than for the document classification task, and we anticipated individual sentences to be more difficult to classify because each sentence only carries a fraction of the information held in an entire document. It is common to have short, simple sentences as part of ordi-

Classifier	All features	Lexical	POS	Surface	Parse
CW	86.40%	92.95%	**95.80%**	69.80%	64.60%
MIRA	**97.50%**	86.55%	94.55%	79.65%	66.90%
MaxEnt	**99.45%**	78.75%	96.25%	86.90%	80.70%
SVM	**99.90%**	75.00%	96.60%	89.75%	82.70%

Table 7: Mean accuracy of all classifiers on the document classification task.

Classifier	All features	POS	Surface	Parse
CW	73.20%	**74.45%**	57.40%	62.25%
MIRA	71.15%	**72.65%**	56.50%	56.45%
MaxEnt	**80.80%**	77.65%	71.30%	69.00%
SVM	**77.00%**	76.40%	72.55%	73.00%

Table 8: Mean accuracy of all classifiers on the sentence classification task.

Category	Mean accuracy change
Everyday life	+1.42%
Geography	-4.29%
History	-1.01%
Literature	-1.84%
Media	-0.56%
People	-0.20%
Religion	-0.56%
Science	-2.50%

Table 9: Mean accuracy drop for a CW model trained on one category and tested on all other categories. Negative numbers indicate a decrease in performance.

nary English text, although they will not make up the whole. However results were still promising, with between 72% and 80% mean accuracy. With CW and MIRA, the classifiers benefited from training on all categories, while MaxEnt and SVM in-category and all-category models achieved similar accuracy levels, but the results on cross-category tests were more variable than in the document classification. There was also no consistency across features and classifiers with regard to category-to-category classification. Overall the results of the sentence classification task are encouraging and show promise for detecting individual simple sentences taken out of context.

6.1 Discussion

The classifiers performed robustly for the document-level classification task, although the corpus itself may have biased the model due to the longer average length of ordinary documents, which we tried to address by filtering out articles with only one or two sentences. Cursory inspection suggests that there is overlap between many Simple Wikipedia articles and their corresponding ordinary English articles, since a large number of Simple Wikipedia documents appear to be generated directly from the English Wikipedia articles with more complicated subsections of the documents omitted from the Simple article.

The sentence classification task could be improved by better labeling of sentences. In these experiments, we assumed that every sentence in an ordinary document would be ordinary (i.e., not simple) and vice versa for simple documents. However it is not the case that ordinary English text contains only complicated sentences. In future research we can use human annotated sentences for building the classifiers. The features we used in this research suggest that simple text is created from categorical lexical and syntactic replacement, but more complicated, technical, or detailed oriented text may require more rewriting, and would be of more interest in future research.

7 Conclusion and Future Work

We have demonstrated the ability to automatically identify texts as either simple or ordinary at both the document and sentence levels using a variety of features based on the word usage and grammatical structures in text. Our statistical analysis has identi-

fied relevant features for this task accessible to computational systems. Immediate applications of the classifiers created in this research for text simplification include editing tools that can identify parts of a text that may be difficult to understand or for word processors, in order to notify writers of complicated sentences in real time.

Using this initial exploration of Simple Wikipedia, we plan to continue working in a number of directions. First, we will explore additional robust indications of text difficulty. For example, Aluísio et al. (2008) claim that sentences that are easier to read are also easier to parse, so the entropy of the parser or confidence in the output may be indicative of a text's difficulty. Additionally, language models trained on large corpora can assign probability scores to texts, which may indicate text difficulty. Of particular interest are syntactic language models that incorporate some of the syntactic observations in this paper (Filimonov and Harper, 2009).

Our next goal will be to look at parallel sentences to learn rules for simplifying text. One of the advantages of the Wikipedia collection is the parallel articles in ordinary English Wikipedia and Simple Wikipedia. While the content of the articles can differ, these are excellent examples of comparable texts that can be useful for learning simplification rules. Such learning can draw from machine translation, which learns rules that translate between languages. The related task of paraphrase extraction could also provide comparable phrases, one of which can be identified as a simplified version of the other (Bannard and Callison-Burch, 2005). An additional resource available in Simple Wikipedia is the flagging of articles as not simple. By examining the revision history of articles whose flags have been changed, we can discover changes that simplified texts. Initial work on this topic has automatically learned which edits correspond to text simplifications (Yatskar et al., 2010).

Text simplification may necessitate the removal of whole phrases, sentences, or even paragraphs, as, according to the writing guidelines for Wikipedia Simple (Wikipedia, 2009), the articles should not exceed a specified length, and some concepts may not be explainable using the lexicon of Basic English. In some situations, adding new text to explain confusing but crucial points may serve to aid the reader, and text generation needs to be further investigated to make text simplification an automatic process.

Acknowledgements

The authors would like to thank Mary Harper for her help in parsing our corpus.

References

S.M. Aluísio, L. Specia, T.A.S. Pardo, E.G. Maziero, and R.P.M. Fortes. 2008. Brazilian portuguese automatic text simplification systems. In *DocEng*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Association for Computational Linguistics (ACL)*.

A.L. Berger, V.J.D. Pietra, and S.A.D. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

F. Biadsy, J. Hirschberg, E. Filatova, and LLC InforSense. 2008. An unsupervised approach to biography production using wikipedia. In *Association for Computational Linguistics (ACL)*.

S. Bird and E. Loper. 2004. Nltk: The natural language toolkit. *Proceedings of the ACL demonstration session*, pages 214–217.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *European Conference for Computational Linguistics (EACL)*, volume 2006, pages 249–256.

Y. Canning, J. Tait, J. Archibald, and R. Crawley. 2000. Cohesive generation of syntactically simplified newspaper text. *Lecture notes in computer science*, pages 145–150.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying text for language-impaired readers. In *European Conference for Computational Linguistics (EACL)*, pages 269–270.

J. Chae and A. Nenkova. 2009. Predicting the fluency of text with shallow structural features. In *European Conference for Computational Linguistics (EACL)*, pages 139–147.

E. Charniak and M. Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Association for Computational Linguistics (ACL)*, page 180. Association for Computational Linguistics.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research (JMLR)*.

W. Daelemans, A. Höthker, and E Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Conference on Language Resources and Evaluation (LREC)*, pages 1045–1048.

Ludovic Denoyer and Patrick Gallinari. 2006. The Wikipedia XML Corpus. *SIGIR Forum*.

S. Devlin and G. Unthank. 2006. Helping aphasic people process online information. In *SIGACCESS Conference on Computers and Accessibility*.

Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. In *International Conference on Machine Learning (ICML)*.

L. Feng, N. Elhadad, and M. Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *European Conference for Computational Linguistics (EACL)*.

Denis Filimonov and Mary Harper. 2009. A joint language model with fine-grain syntactic tags. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Z. Gantner and L. Schmidt-Thieme. 2009. Automatic Content-based Categorization of Wikipedia Articles. In *Association for Computational Linguistics (ACL)*.

Z. Huang and M. Harper. 2009. Self-training pcfg grammars with latent annotations across languages. In *Empirical Methods in Natural Language Processing (EMNLP)*.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*.

T. Kiss and J. Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

R. Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

C. Muller and I. Gurevych. 2008. Using wikipedia and wiktionary in domain-specific information retrieval. In *Working Notes of the Annual CLEF Meeting*. Springer.

C.K. Ogden. 1930. *Basic English: A General Introduction with Rules and Grammar*. Paul Treber & Co., Ltd.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*.

S.E. Petersen and M. Ostendorf. 2007. Text simplification for language learners: A corpus analysis. In *The Speech and Language Technology for Education Workshop*, pages 69–72.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Association for Computational Linguistics (ACL)*.

S.P. Ponzetto and M. Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

R. Schenkel, F. Suchanek, and G. Kasneci. 2007. YAWN: A semantically annotated Wikipedia XML corpus. In *Proceedings of GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW2007)*.

A. Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.

Jason Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Y. Versley, S.P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, and A. Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Association for Computational Linguistics (ACL) Demo Session*.

Voice Of America. 2009. Word book, 2009 edition. www.voaspecialenglish.com, February.

Wikipedia. 2009. Simple wikipedia english. http://en.wikipedia.org/wiki/Citing_Wikipedia, October.

X. Yang and J. Su. 2007. Coreference resolution using semantic relatedness information from automatically discovered patterns. In *Association for Computational Linguistics (ACL)*.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Experiments with unsupervised extraction of lexical simplifications. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Y. Ye, M. Zhou, and C.Y. Lin. 2007. Sentence level machine translation evaluation as a ranking problem: one step aside from BLEU. In *ACL Workshop on statistical machine translation*.