Measuring Contextual Fitness Using Error Contexts Extracted from the Wikipedia Revision History

Torsten Zesch
Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information, Frankfurt

Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt

http://www.ukp.tu-darmstadt.de

Abstract

We evaluate measures of contextual fitness on the task of detecting real-word spelling errors. For that purpose, we extract naturally occurring errors and their contexts from the Wikipedia revision history. We show that such natural errors are better suited for evaluation than the previously used artificially created errors. In particular, the precision of statistical methods has been largely over-estimated, while the precision of knowledge-based approaches has been under-estimated. Additionally, we show that knowledge-based approaches can be improved by using semantic relatedness measures that make use of knowledge beyond classical taxonomic relations. Finally, we show that statistical and knowledge-based methods can be combined for increased performance.

1 Introduction

Measuring the contextual fitness of a term in its context is a key component in different NLP applications like speech recognition (Inkpen and Désilets, 2005), optical character recognition (Wick et al., 2007), co-reference resolution (Bean and Riloff, 2004), or malapropism detection (Bolshakov and Gelbukh, 2003). The main idea is always to test what fits better into the current context: the actual term or a possible replacement that is phonetically, structurally, or semantically similar. We are going to focus on malapropism detection as it allows evaluating measures of contextual fitness in a more direct way than evaluating in a complex application which always entails influence from other components, e.g. the quality of the optical character recognition module (Walker et al., 2010).

A malapropism or real-word spelling error occurs when a word is replaced with another correctly spelled word which does not suit the context, e.g. "People with lots of *honey* usually live in big houses.", where 'money' was replaced with 'honey'. Besides typing mistakes, a major source of such errors is the failed attempt of automatic spelling correctors to correct a misspelled word (Hirst and Budanitsky, 2005). A real-word spelling error is hard to detect, as the erroneous word is not misspelled and fits syntactically into the sentence. Thus, measures of contextual fitness are required to detect words that do not fit their contexts.

Existing measures of contextual fitness can be categorized into knowledge-based (Hirst and Budanitsky, 2005) and statistical methods (Mays et al., 1991; Wilcox-OHearn et al., 2008). Both test the lexical cohesion of a word with its context. For that purpose, knowledge-based approaches employ the structural knowledge encoded in lexical-semantic networks like WordNet (Fellbaum, 1998), while statistical approaches rely on co-occurrence counts collected from large corpora, e.g. the Google Web1T corpus (Brants and Franz, 2006).

So far, evaluation of contextual fitness measures relied on artificial datasets (Mays et al., 1991; Hirst and Budanitsky, 2005) which are created by taking a sentence that is known to be correct, and replacing a word with a similar word from the vocabulary. This has a couple of disadvantages: (i) the replacement might be a synonym of the original word and perfectly valid in the given context, (ii) the generated error might

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 529–538,
Avignon, France, April 23 - 27 2012.

be very unlikely to be made by a human, and (iii) inserting artificial errors often leads to unnatural sentences that are quite easy to correct, e.g. if the word class has changed. However, even if the word class is unchanged, the original word and its replacement might still be variants of the same lemma, e.g. a noun in singular and plural, or a verb in present and past form. This usually leads to a sentence where the error can be easily detected using syntactical or statistical methods, but is almost impossible to detect for knowledge-based measures of contextual fitness, as the meaning of the word stays more or less unchanged. To estimate the impact of this issue, we randomly sampled 1,000 artificially created real-word spelling errors[1] and found 387 singular/plural pairs and 57 pairs which were in another direct relation (e.g. adjective/adverb). This means that almost half of the artificially created errors are not suited for an evaluation targeted at finding optimal measures of contextual fitness, as they over-estimate the performance of statistical measures while underestimating the potential of semantic measures. In order to investigate this issue, we present a framework for mining naturally occurring errors and their contexts from the Wikipedia revision history. We use the resulting English and German datasets to evaluate statistical and knowledge-based measures.

We make the full experimental framework publicly available[2] which will allow reproducing our experiments as well as conducting follow-up experiments. The framework contains (i) methods to extract natural errors from Wikipedia, (ii) reference implementations of the knowledge-based and the statistical methods, and (iii) the evaluation datasets described in this paper.

2 Mining Errors from Wikipedia

Measures of contextual fitness have previously been evaluated using artificially created datasets, as there are very few sources of sentences with naturally occurring errors and their corrections. Recently, the revision history of Wikipedia has been introduced as a valuable knowledge source for NLP (Nelken and Yamangil, 2008; Yatskar et al., 2010). It is also a possible source of natural errors, as it is likely that Wikipedia editors make real-word spelling errors at some point, which are then corrected in subsequent revisions of the same article. The challenge lies in discriminating real-word spelling errors from all sorts of other changes, including non-word spelling errors, reformulations, or the correction of wrong facts. For that purpose, we apply a set of precision-oriented heuristics narrowing down the number of possible error candidates. Such an approach is feasible, as the high number of revisions in Wikipedia allows to be extremely selective.

2.1 Accessing the Revision Data

We access the Wikipedia revision data using the freely available Wikipedia Revision Toolkit (Ferschke et al., 2011) together with the JWPL Wikipedia API (Zesch et al., 2008a).[3] The API outputs plain text converted from Wiki-Markup, but the text still contains a small portion of left-over markup and other artifacts. Thus, we perform additional cleaning steps removing (i) tokens with more than 30 characters (often URLs), (ii) sentences with less than 5 or more than 200 tokens, and (iii) sentences containing a high fraction of special characters like ':' usually indicating Wikipedia-specific artifacts like lists of language links. The remaining sentences are part-of-speech tagged and lemmatized using TreeTagger (Schmid, 2004). Using these cleaned and annotated articles, we form pairs of adjacent article revisions (r_i and r_{i+1}).

2.2 Sentence Alignment

Fully aligning all sentences of the adjacent revisions is a quite costly operation, as sentences can be split, joined, replaced, or moved in the article. However, we are only looking for sentence pairs which are almost identical except for the real-word spelling error and its correction. Thus, we form all sentence pairs and then apply an aggressive but cheap filter that rules out all sentences which (i) are equal, or (ii) whose lengths differ more than a small number of characters. For the resulting much smaller subset of sentence pairs, we compute the Jaro distance (Jaro, 1995) between each pair. If the distance exceeds a certain threshold t_{sim} (0.05 in this case), we do not further consider the pair. The small amount of remaining sentence pairs is passed to the sentence pair filter for in-depth inspection.

[1] The same artificial data as described in Section 3.2.

[2] http://code.google.com/p/dkpro-spelling-asl/

[3] http://code.google.com/p/jwpl/

2.3 Sentence Pair Filtering

The sentence pair filter further reduces the number of remaining sentence pairs by applying a set of heuristics including *surface level* and *semantic level* filters. Surface level filters include:

Replaced Token Sentences need to consist of identical tokens, except for one replaced token.

No Numbers The replaced token may not be a number.

UPPER CASE The replaced token may not be in upper case.

Case Change The change should not only involve case changes, e.g. changing 'english' into 'English'.

Edit Distance The edit distance between the replaced token and its correction need to be below a certain threshold.

After applying the surface level filters, the remaining sentence pairs are well-formed and contain exactly one changed token at the same position in the sentence. However, the change does not need to characterize a real-word spelling error, but could also be a normal spelling error or a semantically motivated change. Thus, we apply a set of semantic filters:

Vocabulary The replaced token needs to occur in the vocabulary. We found that even quite comprehensive word lists discarded too many valid errors as Wikipedia contains articles from a very wide range of domains. Thus, we use a frequency filter based on the Google Web1T n-gram counts (Brants and Franz, 2006). We filter all sentences where the replaced token has a very low unigram count. We experimented with different values and found 25,000 for English and 10,000 for German to yield good results.

Same Lemma The original token and the replaced token may not have the same lemma, e.g. 'car' and 'cars' would not pass this filter.

Stopwords The replaced token should not be in a short list of stopwords (mostly function words).

Named Entity The replaced token should not be part of a named entity. For this purpose, we applied the Stanford NER (Finkel et al., 2005).

Normal Spelling Error We apply the Jazzy spelling detector[4] and rule out all cases in which it is able to detect the error.

Semantic Relation If the original token and the replaced token are in a close lexical-semantic relations, the change is likely to be semantically motivated, e.g. if "house" was replaced with "hut". Thus, we do not consider cases, where we detect a direct semantic relation between the original and the replaced term. For this purpose, we use WordNet (Fellbaum, 1998) for English and GermaNet (Lemnitzer and Kunze, 2002) for German.

[4] http://jazzy.sourceforge.net/

3 Resulting Datasets

3.1 Natural Error Datasets

Using our framework for mining real-word spelling errors in context, we extracted an English dataset[5], and a German dataset[6]. Although the output generally was of high quality, manual post-processing was necessary[7], as (i) for some pairs the available context did not provide enough information to decide which form was correct, and (ii) a problem that might be specific to Wikipedia – vandalism. The revisions are full of cases where words are replaced with similar sounding but greasy alternatives. A relatively mild example is "In romantic comedies, there is a love story about a man and a woman who fall in love, along with silly or funny comedy *farts*.", where 'parts' was replaced with 'farts' only to be changed back shortly afterwards by a Wikipedia vandalism hunter. We removed all cases that resulted from obvious vandalism. For further experiments, a small list of offensive terms could be added to the stopword list to facilitate this process.

A connected problem is correct words that get falsely corrected by Wikipedia editors (without the malicious intend from the previous examples, but with similar consequences). For example, the initially correct sentence "Dung beetles roll it into a ball, sometimes being up to 50 times their own weight." was 'corrected' by exchanging *weight* with *wait*. We manually removed such obvious mistakes, but are still left with some borderline cases. In the sentence "By the 1780s the *goals* of England were so full that convicts were often chained up in rotting old ships." the obvious error

[5] Using a revision dump from April 5, 2011.

[6] Using a revision dump from August 13, 2010.

[7] The most efficient and precise way of finding real-word spelling errors would of course be to apply measures of contextual fitness. However, the resulting dataset would then only contain errors that are detectable by the measures we want to evaluate – a clearly unacceptable bias. Thus, a certain amount of manual validation is inevitable.

'goal' was changed by some Wikipedia editor to 'jail'. However, actually it should have been the old English form for jail 'gaol' which can be deduced when looking at the full context and later versions of the article. We decided to not remove these rare cases, because 'jail' is a valid correction in this context.

After manual inspection, we are left with 466 English and 200 German errors. Given that we restricted our experiment to 5 million English and German revisions, much larger datasets can be extracted if the whole revision history is taken into account. Our snapshot of the English Wikipedia contains $305{\cdot}10^6$ revisions. Even if not all of them correspond to article revisions, it is safe to assume that more than 10,000 real-word spelling errors can be extracted from this version of Wikipedia.

Using the same amount of source revisions, we found significantly more English than German errors. This might be due to (i) English having more short nouns or verbs than German that are more likely to be confused with each other, and (ii) the English Wikipedia being known to attract a larger amount of non-native editors which might lead to higher rates of real-word spelling errors. However, this issue needs to be further investigated e.g. based on comparable corpora build on the basis of different language editions of Wikipedia. Further refining the identification of real-word errors in Wikipedia would allow evaluating how frequent such errors actually occur, and how long it takes the Wikipedia editors to detect them. If errors persist over a long time, using measures of contextual fitness for detection would be even more important.

Another interesting observation is that the average edit distance is around 1.4 for both datasets. This means that a substantial proportion of errors involve more than one edit operation. Given that many measures of contextual fitness allow at most one edit, many naturally occurring errors will not be detected. However, allowing a larger edit distance enormously increases the search space resulting in increased run-time and possibly decreased detection precision due to more false positives.

3.2 Artificial Error Datasets

In contrast to the quite challenging process of mining naturally occurring errors, creating artificial errors is relatively straightforward. From a corpus that is known to be free of spelling errors, sentences are randomly sampled. For each sentence, a random word is selected and all strings with edit distance smaller than a given threshold (2 in our case) are generated. If one of those generated strings is a known word from the vocabulary, it is picked as the artificial error.

Previous work on evaluating real-word spelling correction (Hirst and Budanitsky, 2005; Wilcox-OHearn et al., 2008; Islam and Inkpen, 2009) used a dataset sampled from the Wall Street Journal corpus which is not freely available. Thus, we created a comparable English dataset of 1,000 artificial errors based on the easily available Brown corpus (Francis W. Nelson and Kuçera, 1964).[8] Additionally, we created a German dataset with 1,000 artificial errors based on the TIGER corpus.[9]

4 Measuring Contextual Fitness

There are two main approaches for measuring the contextual fitness of a word in its context: the statistical (Mays et al., 1991) and the knowledge-based approach (Hirst and Budanitsky, 2005).

4.1 Statistical Approach

Mays et al. (1991) introduced an approach based on the noisy-channel model. The model assumes that the correct sentence s is transmitted through a noisy channel adding 'noise' which results in a word w being replaced by an error e leading the wrong sentence s' which we observe. The probability of the correct word w given that we observe the error e can be computed as $P(w|e) = P(w) \cdot P(e|w)$. The channel model $P(e|w)$ describes how likely the typist is to make an error. This is modeled by the parameter α.[10] The remaining probability mass $(1 - \alpha)$ is distributed equally among all words in the vocabulary within an edit distance of 1 ($edits(w)$):

$$P(e|w) = \begin{cases} \alpha & \text{if } e = w \\ (1-\alpha)/|edits(w)| & \text{if } e \neq w \end{cases}$$

The source model $P(w)$ is estimated using a trigram language model, i.e. the probability of the

[8] http://www.archive.org/details/BrownCorpus (CC-by-na).

[9] http://www.ims.uni-stuttgart.de/projekte/TIGER/
The corpus contains 50,000 sentences of German newspaper text, and is freely available under a non-commercial license.

[10] We optimize α on a held-out development set of errors.

intended word w_i is computed as the conditional probability $P(w_i|w_{i-1}w_{i-2})$. Hence, the probability of the correct sentence $s = w_1 \dots w_n$ can be estimated as

$$P(s) = \prod_{i=1}^{n+2} P(w_i|w_{i-1}w_{i-2})$$

The set of candidate sentences S_c contains all versions of the observed sentence s' derived by replacing one word with a word from $edits(w)$, while all other words in the sentence remain unchanged. The correct sentence s is those sentence from S_c that maximizes $P(s|s') = \arg\max_{s \in S_c} P(s) \cdot P(s'|s)$.

4.2 Knowledge Based Approach

Hirst and Budanitsky (2005) introduced a knowledge-based approach that detects real-word spelling errors by checking the semantic relations of a target word with its context. For this purpose, they apply WordNet as the source of lexical-semantic knowledge.

The algorithm flags all words as error candidates and then applies filters to remove those words from further consideration that are unlikely to be errors. First, the algorithm removes all closed-class word candidates as well as candidates which cannot be found in the vocabulary. Candidates are then tested for having lexical cohesion with their context, by (i) checking whether the same surface form or lemma appears again in the context, or (ii) a semantically related concept is found in the context. In both cases, the candidate is removed from the list of candidates. For each remaining possible real-word spelling error, edits are generated by inserting, deleting, or replacing characters up to a certain edit distance (usually 1). Each edit is then tested for lexical cohesion with the context. If at least one of it fits into the context, the candidate is selected as a real-word error.

Hirst and Budanitsky (2005) use two additional filters: First, they remove candidates that are "common non-topical words". It is unclear how the list of such words was compiled. Their list of examples contains words like 'find' or 'world' which we consider to be perfectly valid candidates. Second, they also applied a filter using a list of known multi-words, as the probability for words to accidentally form multi-words is low. It is unclear which list was used. We could use multi-words from WordNet, but coverage would be rather limited. We decided not to use both filters in order to better assess the influence of the underlying semantic relatedness measure on the overall performance.

Dataset	P	R	F
Artificial-English	.77	.50	.60
Natural-English	.54	.26	.35
Artificial-German	.90	.49	.63
Natural-German	.77	.20	.32

Table 1: Performance of the statistical approach using a trigram model based on Google Web1T.

The knowledge based approach uses semantic relatedness measures to determine the cohesion between a candidate and its context. In the experiments by Budanitsky and Hirst (2006), the measure by (Jiang and Conrath, 1997) yields the best results. However, a wide range of other measures have been proposed, cf. (Zesch and Gurevych, 2010). Some measures using a wider definition of semantic relatedness (Gabrilovich and Markovitch, 2007; Zesch et al., 2008b) instead of only using taxonomic relations in a knowledge source.

As semantic relatedness measures usually return a numeric value, we need to determine a threshold θ in order to come up with a binary related/unrelated decision. Budanitsky and Hirst (2006) used a characteristic gap in the standard evaluation dataset by Rubenstein and Goodenough (1965) that separates unrelated from related word pairs. We do not follow this approach, but optimize the threshold on a held-out development set of real-word spelling errors.

5 Results & Discussion

In this section, we report on the results obtained in our evaluation of contextual fitness measures using artificial and natural errors in English and German.

5.1 Statistical Approach

Table 1 summarizes the results obtained by the statistical approach using a trigram model based on the Google Web1T data (Brants and Franz, 2006). On the English artificial errors, we observe a quite high F-measure of .60 that drops to

Dataset	N-gram model	Size	P	R	F
Art-En	Google Web	$7 \cdot 10^{11}$	.77	.50	.60
		$7 \cdot 10^{10}$	.78	.48	.59
		$7 \cdot 10^{9}$	.76	.42	.54
	Wikipedia	$2 \cdot 10^{9}$	.72	.37	.49
Nat-En	Google Web	$7 \cdot 10^{11}$	.54	.26	.35
		$7 \cdot 10^{10}$	.51	.23	.31
		$7 \cdot 10^{9}$	.46	.19	.27
	Wikipedia	$2 \cdot 10^{9}$	.49	.19	.27
Art-De	Google Web	$8 \cdot 10^{10}$	.90	.49	.63
		$8 \cdot 10^{9}$	.90	.47	.61
		$8 \cdot 10^{8}$	.88	.36	.51
	Wikipedia	$7 \cdot 10^{8}$	.90	.37	.52
Nat-De	Google Web	$8 \cdot 10^{10}$	.77	.20	.32
		$8 \cdot 10^{9}$	.68	.14	.23
		$8 \cdot 10^{8}$	.65	.10	.17
	Wikipedia	$7 \cdot 10^{8}$	.70	.13	.22

Table 2: Influence of the n-gram model on the performance of the statistical approach.

.35 when switching to the naturally occurring errors which we extracted from Wikipedia. On the German dataset, we observe almost the same performance drop (from .63 to .32).

These observations correspond to our earlier analysis where we showed that the artificial data contains many cases that are quite easy to correct using a statistical model, e.g. where a plural form of a noun is replaced with its singular form (or vice versa) as in "I bought a car." vs. "I bought a cars.". The naturally occurring errors often contain much harder contexts, as shown in the following example: "Through the open window they heard sounds below in the street: cartwheels, a tired horse's plodding step, vices." where 'vices' should be corrected to 'voices'. While the lemma 'voice' is clearly semantically related to other words in the context like 'hear' or 'sound', the position at the end of the sentence is especially difficult for the trigram-based statistical approach. The only trigram that connects the error to the context is ('step', ',', *vices/voices*) which will probably yield a low frequency count even for very large trigram models. Higher order n-gram models would help, but suffer from the usual data-sparseness problems.

Influence of the N-gram Model For building the trigram model, we used the Google Web1T data, which has some known quality issues and is not targeted towards the Wikipedia articles from which we sampled the natural errors. Thus, we also tested a trigram model based on Wikipedia. However, it is much smaller than the Web model, which leads us to additionally testing smaller Web models. Table 2 summarizes the results.

We observe that "more data is better data" still holds, as the largest Web model always outperforms the Wikipedia model in terms of recall. If we reduce the size of the Web model to the same order of magnitude as the Wikipedia model, the performance of the two models is comparable. We would have expected to see better results for the Wikipedia model in this setting, but its higher quality does not lead to a significant difference.

Even if statistical approaches quite reliably detect real-word spelling errors, the size of the required n-gram models remains a serious obstacle for use in real-world applications. The English Web1T trigram model is about 25GB, which currently is not suited for being applied in settings with limited storage capacities e.g. for intelligent input assistance in mobile devices. As we have seen above, using smaller models will decrease recall to a point where hardly any error will be detected anymore. Thus, we will now have a look on knowledge-based approaches which are less demanding in terms of the required resources.

Dataset	P	R	F
Artificial-English	.26	.15	.19
Natural-English	.29	.18	.23
Artificial-German	.47	.16	.24
Natural-German	.40	.13	.19

Table 3: Performance of the knowledge-based approach using the JiangConrath semantic relatedness measure.

5.2 Knowledge-based Approach

Table 3 shows the results for the knowledge-based measure. In contrast to the statistical approach, the results on the artificial errors are not higher than on the natural errors, but almost equal for German and even lower for English; another piece of evidence supporting our view that the properties of artificial datasets over-estimate the performance of statistical measures.

Influence of the Relatedness Measure As was pointed out before, Budanitsky and Hirst (2006)

Dataset	Measure	θ	P	R	F
Art-En	JiangConrath	0.5	.26	.15	.19
	Lin	0.5	.22	.17	.19
	Lesk	0.5	.19	.16	.17
	ESA-Wikipedia	0.05	**.43**	.13	.20
	ESA-Wiktionary	0.05	.35	**.20**	**.25**
	ESA-Wordnet	0.05	.33	.15	.21
Nat-En	JiangConrath	0.5	.29	.18	.23
	Lin	0.5	.26	**.21**	.23
	Lesk	0.5	.19	.19	.19
	ESA-Wikipedia	0.05	**.48**	.14	.22
	ESA-Wiktionary	0.05	.39	**.21**	**.27**
	ESA-Wordnet	0.05	.36	.15	.21

Table 4: Performance of knowledge-based approach using different relatedness measures.

Dataset	Comb.-Strategy	P	R	F
Artificial-English	Best-Single	.77	.50	.60
	Union	.52	**.55**	.54
	Intersection	**.91**	.15	.25
Natural-English	Best-Single	.54	.26	.35
	Union	.40	**.36**	.38
	Intersection	**.82**	.11	.19

Table 5: Results obtained by a combination of the best statistical and knowledge-based configuration. 'Best-Single' is the best precision or recall obtained by a single measure. 'Union' merges the detections of both approaches. 'Intersection' only detects an error if both methods agree on a detection.

show that the measure by Jiang and Conrath (1997) yields the best results in their experiments on malapropism detection. In addition, we test another path-based measure by Lin (1998), the gloss-based measure by Lesk (1986), and the ESA measure (Gabrilovich and Markovitch, 2007) based on concept vectors from Wikipedia, Wiktionary, and WordNet. Table 4 summarizes the results. In contrast to the findings of Budanitsky and Hirst (2006), JiangConrath is not the best path-based measure, as Lin provides equal or better performance. Even more importantly, other (non path-based) measures yield better performance than both path-based measures. Especially ESA based on Wiktionary provides a good overall performance, while ESA based on Wikipedia provides excellent precision. The advantage of ESA over the other measure types can be explained with its ability to incorporate semantic relationships beyond classical taxonomic relations (as used by path-based measures).

5.3 Combining the Approaches

The statistical and the knowledge-based approach use quite different methods to assess the contextual fitness of a word in its context. This makes it worthwhile trying to combine both approaches. We ran the statistical method (using the full Wikipedia trigram model) and the knowledge-based method (using the ESA-Wiktionary relatedness measure) in parallel and then combined the resulting detections using two strategies: (i) we merge the detections of both approaches in order to obtain higher recall ('Union'), and (ii) we only count an error as detected if both methods agree on a detection ('Intersection'). When comparing the combined results in Table 5 with the best precision or recall obtained by a single measure ('Best-Single'), we observe that precision can be significantly improved using the 'Union' strategy, while recall is only moderately improved using the 'Intersect' strategy. This means that (i) a large subset of errors is detected by both approaches that due to their different sources of knowledge mutually reinforce the detection leading to increased precision, and (ii) a small but otherwise undetectable subset of errors requires considering detections made by one approach only.

6 Related Work

To our knowledge, we are the first to create a dataset of naturally occurring errors based on the revision history of Wikipedia. Max and Wisniewski (2010) used similar techniques to create a dataset of errors from the French Wikipedia. However, they target a wider class of errors including non-word spelling errors, and their class of real-word errors conflates malapropisms as well as other types of changes like reformulations. Thus, their dataset cannot be easily used for our purposes and is only available in French, while our framework allows creating datasets for all major languages with minimal manual effort.

Another possible source of real-word spelling errors are learner corpora (Granger, 2002), e.g. the Cambridge Learner Corpus (Nicholls, 1999). However, annotation of errors is difficult and costly (Rozovskaya and Roth, 2010), only a small fraction of observed errors will be real-word spelling errors, and learners are likely to make dif-

ferent mistakes than proficient language users.

Islam and Inkpen (2009) presented another statistical approach using the Google Web1T data (Brants and Franz, 2006) to create the n-gram model. It slightly outperformed the approach by Mays et al. (1991) when evaluated on a corpus of artificial errors based on the WSJ corpus. However, the results are not directly comparable, as Mays et al. (1991) used a much smaller n-gram model and our results in Section 5.1 show that the size of the n-gram model has a large influence on the results. Eventually, we decided to use the Mays et al. (1991) approach in our study, as it is easier to adapt and augment.

In a re-evaluation of the statistical model by Mays et al. (1991), Wilcox-OHearn et al. (2008) found that it outperformed the knowledge-based method by Hirst and Budanitsky (2005) when evaluated on a corpus of artificial errors based on the WSJ corpus. This is consistent with our findings on the artificial errors based on the Brown corpus, but - as we have seen in the previous section - evaluation on the naturally occurring errors shows a different picture. They also tried to improve the model by permitting multiple corrections and using fixed-length context windows instead of sentences, but obtained discouraging results.

All previously discussed methods are unsupervised in a way that they do not rely on any training data with annotated errors. However, real-word spelling correction has also been tackled by supervised approaches (Golding and Schabes, 1996; Jones and Martin, 1997; Carlson et al., 2001). Those methods rely on predefined *confusion-sets*, i.e. sets of words that are often confounded e.g. {peace, piece} or {weather, whether}. For each set, the methods learn a model of the context in which one or the other alternative is more probable. This yields very high precision, but only for the limited number of previously defined confusion sets. Our framework for extracting natural errors could be used to increase the number of known confusion sets.

7 Conclusions and Future Work

In this paper, we evaluated two main approaches for measuring the contextual fitness of terms: the statistical approach by Mays et al. (1991) and the knowledge-based approach by Hirst and Budanitsky (2005) on the task of detecting real-word spelling errors. For that purpose, we extracted a dataset with naturally occurring errors and their contexts from the Wikipedia revision history. We show that evaluating measures of contextual fitness on this dataset provides a more realistic picture of task performance. In particular, using artificial datasets over-estimates the performance of the statistical approach, while it underestimates the performance of the knowledge-based approach.

We show that n-gram models targeted towards the domain from which the errors are sampled do not improve the performance of the statistical approach if larger n-gram models are available. We further show that the performance of the knowledge-based approach can be improved by using semantic relatedness measures that incorporate knowledge beyond the taxonomic relations in a classical lexical-semantic resource like WordNet. Finally, by combining both approaches, significant increases in precision or recall can be achieved.

In future work, we want to evaluate a wider range of contextual fitness measures, and learn how to combine them using more sophisticated combination strategies. Both - the statistical as well as the knowledge-based approach - will benefit from a better model of the typist, as not all edit operations are equally likely (Kernighan et al., 1990). On the side of the error extraction, we are going to further improve the extraction process by incorporating more knowledge about the revisions. For example, vandalism is often reverted very quickly, which can be detected when looking at the full set of revisions of an article.

We hope that making the experimental framework publicly available will foster future research in this field, as our results on the natural errors show that the problem is still quite challenging.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We Andreas Kellner and Tristan Miller for checking the datasets, and the anonymous reviewers for their helpful feedback.

References

David Bean and Ellen Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. In *Proc. of HLT/NAACL*, pages 297–304.

Igor A. Bolshakov and Alexander Gelbukh. 2003. On Detection of Malapropisms by Multistage Collocation Testing. In *Proceedings of NLDB-2003, 8th International Workshop on Applications of Natural Language to Information Systems*, number Cic.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Andrew J Carlson, Jeffrey Rosen, and Dan Roth. 2001. Scaling Up Context-Sensitive Text Correction. In *Proceedings of IAAI*.

C Fellbaum. 1998. *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. 2011. Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, pages 97–102, Portland, OR, USA.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, pages 363–370, Morristown, NJ, USA. Association for Computational Linguistics.

Francis W. Nelson and Henry Kuçera. 1964. Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.

Andrew R. Golding and Yves Schabes. 1996. Combining Trigram-based and feature-based methods for context-sensitive spelling correction. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics* -, pages 71–78, Morristown, NJ, USA. Association for Computational Linguistics.

Sylviane Granger, 2002. *A birds-eye view of learner corpus research*, pages 3–33. John Benjamins Publishing Company.

Graeme Hirst and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111, March.

Diana Inkpen and Alain Désilets. 2005. Semantic similarity for detecting recognition errors in automatic speech transcripts. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, number October, pages 49–56, Morristown, NJ, USA. Association for Computational Linguistics.

Aminul Islam and Diana Inkpen. 2009. Real-word spelling correction using Google Web IT 3-grams. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 3 - EMNLP '09*, Morristown, NJ, USA. Association for Computational Linguistics.

M A Jaro. 1995. Probabilistic linkage of large public health data file. *Statistics in Medicine*, 14:491–498.

Jay J Jiang and David W Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics*, Taipei, Taiwan.

Michael P Jones and James H Martin. 1997. Contextual spelling correction using latent semantic analysis. In *Proceedings of the fifth conference on Applied natural language processing* -, pages 166–173, Morristown, NJ, USA. Association for Computational Linguistics.

Mark D Kernighan, Kenneth W Church, and William A Gale. 1990. A Spelling Correction Program Based on a Noisy Channel Model. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 205–210, Helsinki, Finland.

Lothar Lemnitzer and Claudia Kunze. 2002. GermaNet - Representation, Visualization, Application. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 1485–1491.

M Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference*, pages 24–26.

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. In *Proceedings of International Conference on Machine Learning*, pages 296–304, Madison, Wisconsin.

Aurelien Max and Guillaume Wisniewski. 2010. Mining Naturally-occurring Corrections and Paraphrases from Wikipedias Revision History. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3143–3148.

Eric Mays, Fred. J Damerau, and Robert L Mercer. 1991. Context based spelling correction. *Information Processing & Management*, 27(5):517–522.

Rani Nelken and Elif Yamangil. 2008. Mining Wikipedia's Article Revision History for Training Computational Linguistics Algorithms. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (WikiAI)*, WikiAI08.

Diane Nicholls. 1999. The Cambridge Learner Corpus - Error Coding and Analysis for Lexicography and ELT. In *Summer Workshop on Learner Corpora*, Tokyo, Japan.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL Errors: Challenges and Rewards. In *The 5th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL-HLT)*.

H Rubenstein and J B Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

Daniel D. Walker, William B. Lund, and Eric K. Ringger. 2010. Evaluating Models of Latent Document Semantics in the Presence of OCR Errors. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (October):240–250.

M. Wick, M. Ross, and E. Learned-Miller. 2007. Context-sensitive error correction: Using topic models to improve OCR. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, pages 1168–1172. Ieee, September.

Amber Wilcox-OHearn, Graeme Hirst, and Alexander Budanitsky. 2008. Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing (CICLing)*.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368.

Torsten Zesch and Iryna Gurevych. 2010. Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering*, 16(1):25–59.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008a. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008b. Using wiktionary for computing semantic relatedness. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pages 861–867, Chicago, IL, USA, Jul.