Helping Our Own 2011: UKP Lab System Description

Torsten Zesch
Ubiquitous Knowledge Processing (UKP) Lab
Technische Universitt Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract

This paper describes the UKP Lab system participating in the Helping Our Own Challenge 2011. We focus on the correction of real-word spelling errors (RWSEs) that are especially hard to detect. Our highly flexible system architecture is based on UIMA (Ferrucci and Lally, 2004) and integrates state-of-the-art approaches for detecting RWSEs.

1 Introduction

Real-word spelling errors (RWSEs) occur when a word is replaced with another correctly spelled word which is not intended in that context. For example, file '0046' from the development data contains "... untagged *copra* are often used to do emotion classification research.", where the writer mistakenly replaced 'corpora' with 'copra'. As 'copra' (dried coconut meat) is a valid word, the error cannot be detected using a lexicon-based spell checker. In this case, the correction would rather be "... untagged copra *is* often used ..." because of the number agreement error. Real-word spelling errors like "copra/corpora" can only be detected using methods that analyze the context fitness of each term in a sentence.

The example above is tagged with the error class "S" together with other forms of spelling errors. The development data contains relatively few errors in this class, and only a the smaller part of them are RWSEs. However, RWSEs still pose a serious problem, as they give a sentence an unintended meaning which might heavily confuse the reader.

2 System Description

We implemented a general framework for error detection based on the open-source DKPro framework.[1] DKPro is a collection of software components for natural language processing based on the Apache UIMA framework (Ferrucci and Lally, 2004). It comes with a collection of ready-made modules which can be combined to form more complex applications.

Jazzy DKPro already provides a wrapper for the open-source spell checker Jazzy.[2]. Although it is not targeted towards RWSEs, we use it for reasons of comparison with other approaches.

Detecting RWSEs We re-implemented two state-of-the-art approaches: the knowledge-based approach by Hirst and Budanitsky (2005) (**BH2005**) and the statistical approach by Mays et al. (1991) (**MDM1991**). Both approaches test the lexical cohesion of a word with its context.

For that purpose, BH2005 computes the semantic relatedness of a target word with all other words in a certain context window to test whether the target word fits its context. Following Hirst and Budanitsky (2005), we use the semantic relatedness measure by Jiang and Conrath (1997) and WordNet (Fellbaum, 1998) as a knowledge source. If a target word does not fit its context, it is flagged as a possible error. Then, the set of valid words with low edit distance to the target word is computed. Each of the words in this set, that better fits into the given context than the target word, is selected as a possible correction.

[1] http://code.google.com/p/dkpro-core-asl/

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

Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 260–262,
Nancy, France, September 2011. ©2011 Association for Computational Linguistics

	Detection			Recognition			Correction		
Dataset	P	R	S	P	R	S	P	R	S
Jazzy	0.054	0.115	0.073	0.028	0.064	0.039	0.007	0.015	0.009
HB2005	0.093	0.028	0.043	0.048	0.013	0.020	0.009	0.002	0.003
MDM1991 (Google)	0.211	0.026	0.046	0.157	0.020	0.035	0.114	0.015	0.026
MDM1991 (ACL)	0.717	0.004	0.009	0.450	0.003	0.006	**0.450**	0.003	0.006
JoinRWSE	0.095	0.030	0.045	0.055	0.015	0.023	0.020	0.004	0.007
JoinAll	0.051	**0.136**	0.075	0.029	**0.073**	0.041	0.007	**0.016**	0.010
IntersectAll	**1.000**	0.006	0.013	**0.625**	0.004	0.009	0.313	0.003	0.005

Table 1: Overview of evaluation results. Best values are in bold.

The statistical approach (MDM1991) is based on the noisy-channel model assuming 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. Hence, the probability of the correct word w, given the error e is observed, can be computed using a n-gram language model and a model of how likely the typist is to make a certain error. We use two language models: (i) based on the Google Web1T n-gram data (Brants and Franz, 2006), and (ii) based on all the papers in the ACL Anthology Reference Corpus (Bird et al., 2008).

2.1 Combined Approaches

Our framework allows to easily combine spell checkers. In all the combination experiments, we used the MDM1991 with the Google n-gram model.

JoinRWSE Only the two approaches targeted towards RWSEs (i.e. BH2005 and MDM1991) are combined.

JoinAll All three spell checkers (Jazzy, BH2005, and MDM1991) are run in parallel and detections are joined as if only a single spell checker would have been used.

IntersectAll All three spell checkers (Jazzy, BH2005, and MDM1991) are run in parallel, but only errors that are detected by each of the spell checkers are retained.

3 Preliminary Results

As by the time of writing the final results are not yet available, we can only report preliminary results and analyses. Table 1 summarizes the results.

The knowledge-based approach (HB2005) does not perform well, as the documents contain a large amount of domain-specific vocabulary that is either not found in WordNet at all or not with the correct sense. The statistical approach (MDM1991) using the Google n-gram model yields a detection precision of $.21$ which translates into a still acceptable rate of false alarms, but the recall is very low. The detection precision of MDM1991 gets a significant boost using the ACL corpus n-gram model ($P = .72$), but at the price of an even lower recall. However, unlike the other models, MDM1991 with the ACL n-gram model is also able to provide quite good corrections ($P = .45$).

Regarding the combination experiments, we find that joining the two approaches for detecting RWSEs did not significantly increase recall indicating that both approaches more or the less detect the same errors. In contrast, recall significantly increases when joining all approaches which shows that the errors detected by Jazzy are largely complimentary to those detected by the two RWSE approaches.

The “join” combination strategy focuses on recall, but in the setting of this challenge high precision is more important than high recall, as writers might be tempted to take the detected errors and suggested corrections for granted. The result could be a document with more errors than before. Thus, we also used the “intersection“ strategy which should yield better precision. When intersecting the results of all approaches, we obtain perfect precision, but very low detection recall ($.06\%$ translating into 8 overall detections).

When looking at the detected errors by type, we find that MDM1991 (with Google N-grams) detects

50% of all errors in the “S” class. However, to our surprise, it also detects 83% of errors in the “CN” class. Further analyses are necessary to investigate this behavior.

References

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *In Proceedings of Language Resources and Evaluation Conference (LREC 08). Marrakesh, Morocco.*

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia.

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

David Ferrucci and Adam Lally. 2004. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3-4):327–348.

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

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.

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