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

Torsten Zesch

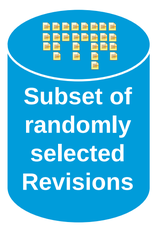
http://www.ukp.tu-darmstadt.de

Motivation

- Between 25% and 40% of spelling errors are valid English words (Kukich, 1992)
- Especially frequent for non-native speakers (Atwell, 1987)
- Often introduced by failed attempt of automatic spelling correctors to correct a misspelled word (Hirst and Budanitsky, 2005).

- Spelling correctors are evaluated using artificially created datasets
- Problems with artificial errors: (i) different word class (This is an/ant example); (ii) not an error (Do you like this color/colour.); (iii) variants of the same lemma (He bought the house/houses).
- Random sample of 1,000 artificially created errors: 387 singular/plural pairs and 57 pairs which were in another direct relation (e.g. adjective/adverb).

Error Mining



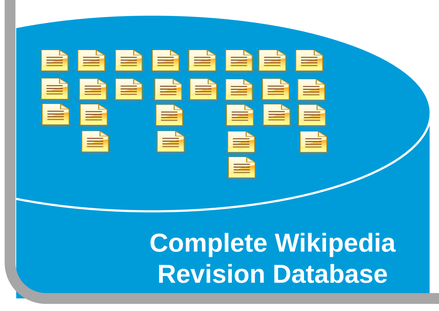
- From two adjacent revisions, create aligned sentence pairs where exactly one token is different.

```
People with lots of honey usually live in big houses.

People with lots of money usually live in big houses.
```

Minimal manual post-processing

Dataset of naturally occurring errors in context



- English Wikipedia Snapshot contains $3 \cdot 10^8$ revisions
- Access to Wikipedia Revision via the JWPL Revision Toolkit (Ferschke, Zesch & Gurevych, 2011 ACL System Demonstration)

Surface Level Sentence Filters

- Replaced token may not be a number
- Replaced token may not be in upper case
- The change should not only involve case changes, e.g. changing 'english' into 'English'
- The edit distance between the replaced token and its correction need to be below a certain threshold

Semantic Filters

- Replaced token needs to be in the vocabulary
- Original token and replaced token may not have the same lemma
- Replaced token must not be a stopword
- Replaced token must not be part of a named entity
- Replaced token must not be detectable using a normal spelling detector (Jazzy)
- 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" (we use Wordnet & GermaNet)..

Experiments

Measuring the Lexical Cohesion of a Word with its Context

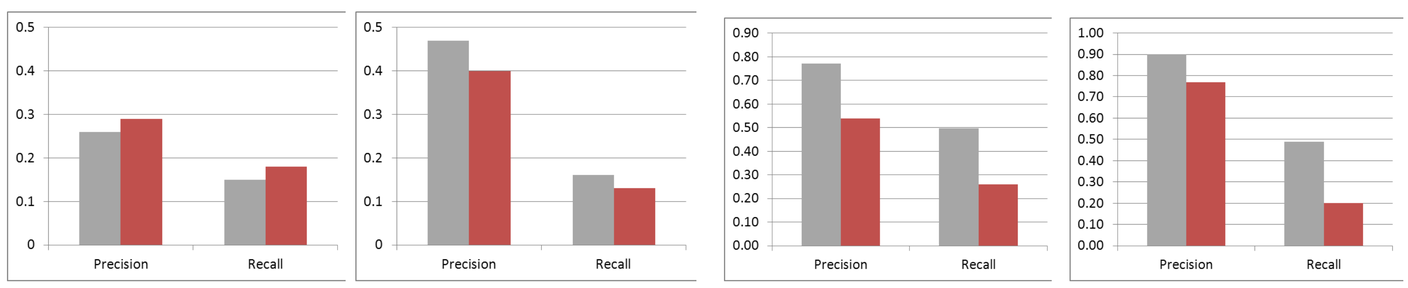
Knowledge-based Approach

- Approach by (Hirst and Budanitsky, 2005)
- Computes the semantic relatedness of a target word with all other words in its context
- Tested a wide range of relatedness measures
- If a target word does not fit its context, it is flagged as a possible error
- If a word with low edit distance to a flagged target word fits better into the given context, it is selected as a possible correction.

Statistical Approach

- Statistical approach by (Mays et al. 1991)
- Based on noisy-channel model
- The probability of the correct word w, given the error e is observed, can be computed using an n-gram language model and a model of how likely the typist is to make a certain error.
- N-gram models based on Google Web1T data (Brants and Franz, 2006) and Wikipedia

English | German | English | German



Artificial Errors

Natural Errors

Influence of Relatedness Measure

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

Influence of N-gram Model Size

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

Conclusions

- Datasets for English and German freely available (mined from 5M revisions, 486 English errors, 200 German errors) – other languages possible
- Much larger datasets can be easily mined from Wikipedia (conservative estimate: about 10,000 errors from complete revision history)
- Revisions generally are a rich source for phenomena, e.g. rephrasing
- Using artificial evaluation datasets over-estimates the performance of the statistical approach, while underestimating the performance of the knowledge-based approach
- Quite large impact of semantic relatedness measures on the knowledge-based approach

References

- Brants T. and A. Franz. 2006. Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia.
- Ferschke O., T. Zesch and I. Gurevych. 2011. Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History. In: Proceedings of the 49th Annual Meeting of the ACL. System Demonstrations, p. 97-102, June 2011.
- Hirst G. and A. Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering, 11(1):87–111.
- Kukich, K. 1992. Techniques for Automatically Correcting Words in Text. ACM Computing Surveys, 24(4), 377-439. New York, NY, USA: ACM.
- Mays E., F. J. Damerau, and R. L. Mercer. 1991. Context based spelling correction. Information Processing & Management, 27(5):517–522.

VolkswagenStiftung

DIPF Bildungsforschung und Bildungsinformation