Helping Our Own 2011: UKP Lab System Description

TECHNISCHE UNIVERSITÄT DARMSTADT

UBIQUITOUS KNOWLEDGE PROCESSING

Torsten Zesch

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

Setup

Traditional Spell Checker

- Jazzy spell checker (http://jazzy.sourceforge.net/)
- As provided in DKPro (http://code.google.com/p/dkpro-core-asl/)

The Java Open Source Spell Checker

Run 1

Measuring Contextual Fitness

- Test the lexical cohesion of a word with its context

Knowledge-based

- Approach by Hirst and Budanitsky '05 (HB2005)
- Computes the semantic relatedness of a target word with all other words in its context
- Jiang and Conrath (1997) Semantic relatedness measure with WordNet
- 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.

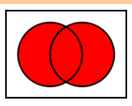
Run 2

nGram-based

- Statistical approach by Mays et al. '91 (MDM1991)
- Based on noisy-channel model
- 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.
- nGram models based on:
  1. Google Web1T n-gram data (Brants and Franz, 2006) — Run 3
  2. ACLAnthology Reference Corpus (Bird et al., 2008) — Run 4

Combination of Approaches

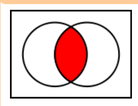
Join



JoinAll: Run 5 = 1 2 3 4

JoinContextFitness: Run 7 = 2 3

Intersection



IntersectAll: Run 6 = 1 2 3 4

Results

Results over all error classes (macro-average on document basis)

			Detection			Recognition			Correction		
	Name	RunId	P	R	S	P	R	S	P	R	S
Single	Jazzy	1	0.054	0.115	0.073	0.028	0.064	0.039	0.007	0.015	0.009
	HB2005	2	0.093	0.028	0.043	0.048	0.013	0.020	0.009	0.002	0.003
	MDM1991 (Google)	3	0.211	0.026	0.046	0.157	0.020	0.035	0.114	0.015	0.026
	MDM1991 (ACL)	4	0.717	0.004	0.009	0.450	0.003	0.006	**0.450**	0.003	0.006
Combined	JoinAll	5	0.051	**0.136**	0.075	0.029	**0.073**	0.041	0.007	**0.016**	0.010
	IntersectAll	6	**1.000**	0.006	0.013	**0.625**	0.004	0.009	0.313	0.003	0.005
	JoinContextFitness	7	0.095	0.030	0.045	0.055	0.015	0.023	0.020	0.004	0.007

Detection results by error class

Class	# errors	Rank	RunID
Article errors	260	5	5
Punctuation errors	206	1	5
Preposition errors	121	4	5
Noun errors	113	1	1
Verb errors	108	1	5
Compound Change errors	66	1	5
Adjective errors	34	1	5
Adverb errors	28	1	5
Conjunction errors	20	3	-
Anaphor errors	14	1	2
Spelling errors	9	3	2
Quantifier errors	7	2	2
Other errors	80	1	5

Conclusions

- Best participating system for 8 out of 13 error classes
- Contextual fitness measures proved generally effective also for error classes not directly targeted
- Combine methods specialized on certain error classes, as there seems to be no "one fits all" approach
- Automated writing assistance stays a challenging task → we only made the first steps to really "helping our own"

References

- Steven Bird et al. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. 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.
- Graeme Hirst and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering, 11(1):87–111.
- 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.

VolkswagenStiftung