AHA: Anagram Hashing Application

Reynaert, Martin

Published in:
Proceedings of CLARIN Annual Conference 2016

Document version:
Early version, also known as pre-print

Publication date:
2016

Link to publication

Citation for published version (APA):
Reynaert, M. (2016). AHA: Anagram Hashing Application. In L. Borin (Ed.), *Proceedings of CLARIN Annual Conference 2016* (pp. 75-79). CLARIN ERIC.

AHA: Anagram Hashing Application

Martin Reynaert
TiCC / Tilburg University
CLST / Radboud University Nijmegen
reynaert@uvt.nl

Abstract

We briefly present AHA, the Anagram Hashing Application, a new web application and service that allows researchers to effortlessly analyse the lexical variation present in their Gold Standard data and to publish the results.

1 Introduction

We present a new web application and service that enables researchers to obtain character confusion lists and attendant frequency statistics from word lists. The tool serves to obtain derived information for a variety of purposes such as diachronical or dialectical language variance studies, spelling and OCR-error profiling, etc.

A major aim of the web application/service provided is to enable researchers to give a succinct quantitative summary of their gold standard data in terms of the commonly accepted Damerau-Levenshtein error categories of insertion, deletion, substitution, transposition and combinations or special variants e.g. regarding capitalization or spacing. Given for instance a corpus-derived list of the words in a diachronical corpus and a contemporary word list, the tool allows for efficiently determining which historical spelling variations are prevalent in the data.

Already in the early days of spelling correction research (Pollock and Zamora, 1983) set a fine example of collecting and analyzing 50,000 English misspellings[1]. These days, very few researchers provide actual analyses of the error types in their Gold Standards. The tool we provide may help to remedy that.

This service is also intended to help researchers in error correction to better come to grips with the problems actually posed by their own test sets. Too often, researchers do not seem to have a clear overview of the challenges present in their own Gold Standard. To exemplify this, we analyse the test set employed for Aspell and compare its statistics to those of the list of common misspellings collected by Wikipedia.

Although we think that what we describe here already presents a useful tool, we intend to extend it with a fuller range of diverse test sets exemplifying a far wider range of problems it can be fruitfully employed to explore in far more depth.

2 Anagram Hashing Application

2.1 AHA: system overview

AHA is a subsystem of the corpus building pipeline PICCL, the acronym for 'Philosophical Integrator of Computational and Corpus Libraries', we are currently developing in the Dutch CLARIN project CLARIAH. We give a brief overview of PICCL in (Reynaert et al., 2015).

Part of the system provided is based on the anagram hashing approach to lexical variation as we first proposed in (Reynaert, 2004). We have recently made available a new implementation of the system we currently call Text-Induced Corpus Clean-up or TICCL as open source software through GitHub[2] (Reynaert, 2016). We enlist our anagram hashing approach to exhaustively charting lexical variation up

[1] Sadly, this valuable resource was lost 'in the mists of time' (Zamora, personal communication).

[2] https://github.com/martinreynaert/TICCL

to a specified Levenshtein Distance here. We further extend it with an available Perl module that rewrites the actual pairs of incorrect and correct words listed in a spelling correction Gold Standard into patterns. These patterns describe actual character matches between word strings as ‘M’ – per character, as ‘D’ for deleted characters, as ‘I’ for inserted ones, ‘S’ for substituted characters and ‘T’ for transposed ones. This then allows for accurately counting the lexical confusions and their combinations that are apparent in one’s Gold Standard and for producing comprehensive summaries of them in a handy table-format.

The patterns returned by the Perl module thus allow us to sufficiently accurately determine, given an incorrect or divergent input string and its aligned perceived correct counterpart, what category of character confusion the pair represents. On the basis of their anagram value difference, using the output of the TICCL modules, we can then retrieve the actual character confusion displayed by the pair.

2.2 The Brew module

The single program we know to be available for describing the patterns displayed by typos is the Perl module Text::Brew[3]. This returns not only the edit distance between a word pair just as e.g. a typo and its orthographically correct word form, but also the pattern of where the two word strings actually diverge and whether this is due to deletions, insertions or substitutions and possible combinations thereof. We in effect abbreviate the module’s output and extend it with patterns for transpositions. We further call on TICCL’s help to help name the patterns seen, i.e. to actually state that for instance these n characters were in fact replaced by these m characters. The latter facility is nevertheless currently limited to the ’reach’ in terms of Levenshtein distance or LD (Levenshtein, 1966) allowed for by the TICCL character confusion list used.

3 Using AHA

We take as our first example data the Aspell evaluation list for English available online.[4] This list provides 547 typos paired to their correct word form. The description states: ”It tries to focus on really bad misspellings”. Our second set of Gold Standard data is the list of 4,509 common misspellings of English words provided by Wikipedia.[5]

Table 1 and Table 2 serve as illustrations of some of the useful output provided by AHA. AHA formats the tables in Latex style, ready for incorporation in one’s paper. The tables were copied and pasted from the output as they are, with the exception of the captions and reference labels, which we needed to yet fill in.

Category	Levenshtein Distance						Total	%
	1	2	3	4	5	6		
deletion	78	14	4				96	17.55
insertion	69	9	1				79	14.44
substitution	108	29	5	5	2		149	27.24
transposition		28					28	5.12
multisingle		23	12	5			40	7.31
multiple		65	42	19	9	3	138	25.23
space deletion	13						13	2.38
space insertion								0.00
capitalisation								0.00
TOTAL	268	168	64	29	11	3	547	
%	48.99	30.71	11.70	5.30	2.01	0.55		99.3

Table 1: Statistics of the error categories in the Aspell 547 items long typo/correction list.

[3] http://search.cpan.org/~kcivey/Text-Brew-0.02/lib/Text/Brew.pm

[4] http://aspell.net/test/cur/batch0.tab

[5] https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings

Category	Levenshtein Distance											Total	%
	1	2	3	4	5	6	7	8	9	10	11		
deletion	1515	74	2									1591	35.28
insertion	808	49	4									861	19.10
substitution	892	124	9		1							1026	22.75
transposition		558	1	3								562	12.46
multisingle		58	10	4		3						75	1.66
multiple		270	69	18	4	1	1	1			1	365	8.09
space deletion	27	2										29	0.64
space insertion													0.00
capitalisation													0.00
TOTAL	3242	1135	95	25	5	4	1	1			1	4509	
%	71.90	25.17	2.11	0.55	0.11	0.09	0.02	0.02			0.02		100.0

Table 2: Statistics of the error categories in the Wikipedia 4,509 items long typo/correction list.

Apart from the table summarising the error or character confusion types observed in the data provided to the system, it naturally also outputs a list of the actual confusion patterns observed and their frequencies.

We observe that the Aspell list has only 79.7% of LD 1 and 2 errors, while in the Wikipedia list these amount to 97,07%. This wide divergence of results obtained from both Gold Standard lists calls for some further interpretation. The Wikipedia list does have some higher LD cases than the Aspell list. This, however, is manageable in so far that Wikipedia goes in for 'absolute correction' as first defined by (Pollock and Zamora, 1984): given a known non-word, replace it by its correct form. Aspell is different in this respect in that in fact it seeks to actually transform the erroneous form and so to arrive at the correct one.

Given the output of AHA for both 'gold standards' we can now easily compare them – not only quantitatively, but also more qualitatively. We contrastively list the top 20 most serious elevated LD cases from the Aspell and Wikipedia collections in Table 3. We see that the Aspell collection is a compendium of pathological spellings. We think one cannot seriously expect a spelling correction system to actually correct these elevated LD cases. (Choudhury et al., 2007) show quite convincingly that beyond LD 4 spelling correction becomes a near impossibility. The Wikipedia list offers more what may be labeled cognitive errors, the result of actual lack of mastery of the English language, exemplified for instance by the overgeneration of weak verbal forms. The worst LD cases are more likely the result of the Wikipedia community disfavouring particular words. We also see some fascinating mistypings, where most all the letters constituting the word intended are actually produced, but none are in their right place.

Given that the Aspell list displays the following pairs: islams > muslims, isreal > israel, johhn > john, judgement > judgment, kippur > kipper, one may very well question the resolutions of these 'typos'. These may well have been the result of a writing task taken by an inept speller. Given that 'isreal' has been resolved as an uncapitalized 'Israel' instead of 'is real' and the presence of te word 'judgment', should perhaps 'johhn kippur' not have been resolved as 'Yom Kippur'? In the absence of the actual text, without contextual clues, one cannot know. The same goes for a great many of the actual pairs in the Aspell list. The typo 'instulation', resolved as 'installation', has only an LD of 1 to 'insulation'. The string 'leasve', resolved as 'leave', may well have had to read 'lease'. We advise to treat the Aspell list with great caution.

4 Future developments

In future work we would like to achieve a kind of laboratory where various approaches to solving lexical variation problems might be easily, fruitfully and honestly compared and evaluated. Given the current tool

Aspell	Wikipedia
psychologist > sicolagest > 6	discomfort > unconfortability > 11
miscellaneous > misilous > 6	muslims > mohammedans > 8
environmental > invermeantial > 6	muslim > muhammadan > 7
Unfortunately > Unformanlly > 5	Premonstratensians > Premonasterians > 6
theoretical > theridically > 5	geometers > geometricians > 6
righten > writeen > 5	geometer > geometrician > 6
Occasionally > Accosinly > 5	1990s > ninties > 6
hyphen > hifine > 5	transcendental > transcendentational > 5
frustrating > frustartaion > 5	taught > teached > 5
environmental > envireminakl > 5	sought > seeked > 5
cockamamie > cocamena > 5	memorable > rememberable > 5
automatically > autoamlly > 5	characteristics > charistics > 5
architecture > aricticure > 5	years > eyasr > 4
architecture > aratictature > 5	would > owudl > 4
Unfortunately > Unfortally > 4	which > hwihc > 4
unconstitutional > unconisitional > 4	Wednesday > wendsay > 4
unconscious > unconscience > 4	think > htikn > 4
tough > taff > 4	subpoena > sepina > 4
theoretical > teridical > 4	strongest > stornegst > 4
substitutions > subisitions > 4	shouldn't > shoudln > 4

Table 3: Qualitative comparison of the top 20 LD cases as present in the Aspell and Wikipedia 'gold standards'.

we know to have only scratched the surface of what we can do. A comprehensive description of one's Gold Standard is very closely linked to actually measuring what one is doing. We therefore envisage further complementing this tool with actual evaluation modules in order to e.g. accurately measure the differences in performance reached when one compares one particular system's output to another's.

By the time of the CLARIN Annual Conference 2016, the system is to be available to the wider community through a recognised CLARIN Centre[6], as well as at Tilburg University.[7]

5 Conclusion

We have given a brief introduction to AHA, the Anagram Hashing Application.

While English is the current Lingua Franca across the world, our tool should be applicable to at least most languages that rely on an alphabetic script for writing. To appeal to a wide audience, we have here applied the tool to two English but divergent lists of typographical errors available online.

On the basis of the statistics derived from typos paired to their correct English word forms we have illustrated what can be learned from the results and demonstrated that this constitutes valuable information which researchers may want to provide in their research reports and papers. We have further explained that the tool, further equipped with modules for actual evaluation, may serve as a laboratory and test bed for further extensions to or refinements of spelling and OCR post-correction systems.

What so far we have done here is to put at researchers' disposal a quick and handy tool to derive useful and comparable statistics from their own Gold Standards for lexical variation in their own languages. We think this application serves a purpose and hope it will find favour widely.

Acknowledgments

Martin Reynaert acknowledges being funded by the Dutch national CLARIN programme CLARIAH Work Packages 2 and 3 in project PICCL and by NWO in project Nederlab.

[6] http://ticclops.clarin.inl.nl/AHA/

[7] http://ticclops.uvt.nl/AHA/

References

[Choudhury et al.2007] Monojit Choudhury, Markose Thomas, Animesh Mukherjee, Anupam Basu, and Niloy Ganguly. 2007. How difficult is it to develop a perfect spell-checker? A cross-linguistic analysis through complex network approach. *Proceedings of the second workshop on TextGraphs: Graph-based algorithms for natural language processing*, pages 81–88.

[Levenshtein1966] V. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Sovjet Physics Doklady*, 10:707–710.

[Pollock and Zamora1983] J.J. Pollock and A. Zamora. 1983. Collection and characterization of spelling errors in scientific and scholarly text. *Journal of the American Society for Information Science*, 34(1):51–58, January.

[Pollock and Zamora1984] J.J. Pollock and A. Zamora. 1984. Automatic spelling correction in scientific and scholarly text. *Commun. ACM*, 27(4):358–368.

[Reynaert et al.2015] Martin Reynaert, Maarten van Gompel, Ko van der Sloot, and Antal van den Bosch. 2015. PICCL: Philosophical Integrator of Computational and Corpus Libraries. In *Proceedings of CLARIN Annual Conference 2015 – Book of Abstracts*, pages 75–79, Wrocław, Poland. CLARIN ERIC.

[Reynaert2004] M. Reynaert. 2004. Text-induced spelling correction. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

[Reynaert2016] Martin Reynaert. 2016. OCR Post-Correction Evaluation of Early Dutch Books Online - Revisited. In Nicoletta Calzolari et a., editor, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).