Evaluation of String Normalisation Modules for String-based Biomedical Vocabularies Alignment with AnAGram

Anique van Berne, A.vanBerne@Elsevier.com Elsevier BV

Veronique Malaisé V.Malaise@Elsevier.com Elsevier BV

Abstract: We evaluate the precision and recall of the different normalization modules of AnAGram: a modular string-based vocabulary alignment tool we built for biomedical vocabularies. The main feature of AnAGram is a targeted transformation using a dictionary of adjective/noun correspondences, which gives interesting results. We find that the classic Porter stemming algorithm needs adaption to the biomedical domain in order to produce quality results.

1. Introduction: AnAGram and Related Work

This paper stems from a product interoperability effort in the biomedical domain through taxonomy alignment. Though requiring a generic tool, each individual alignment requires specific conditions to be optimal, due to lexical idiosyncrasies. AnAGram is constructed as a modular, step-wise, string-based alignment tool (as string-based tools perform well on the anatomical datasets of the OAEI campaign[1]).

AnAGram is built for a local system[2], using hash-table lookup for performance. Matching is modular: a user selects one or multiple modules for processing the source taxonomy. The alignment stops at the first match in the target taxonomy. The modules are ordered to produce results of increasing distance from the original string (similar to a confidence value) and include: exact match; stop word removal (using an independent fine-tuned list); re-ordering (sorting tokens alphabetically for multi-word terms match); stemming (with Porter stemmer[3]); normalization (of non-alpha-numeric characters); substitution (replacing adjective/noun from our substitution dictionary).

The modules correspond to the list by Cheatham and Hitzler[4] of syntactic linguistic processes used by at least one alignment tool in the Ontology Alignment Evaluation Initiative (OAEI)[5]. Chua and Kim's[6] approach is closest to AnAGram, using WordNet[7] for building adjective/noun pairs to improve their matches, where ours is built on the biomedical reference Dorland's (creating a larger substitution dictionary).

---

[1] http://oaei.ontologymatching.org/2013/anatomy/index.html

[2] Dell™ Precision™ T7500, 2x Intel® Xeon® CPU E5620 2.4 GHz processors, 64 GB RAM. Software: Windows 7 Professional 64 bit, Service Pack 1; Perl v5.16.3

[3] http://tartarus.org/martin/PorterStemmer/

[4] http://disi.unitn.it/~p2p/RelatedWork/Matching/strings-iswc13.pdf

[5] http://oaei.ontologymatching.org/2014/

[6] http://www.ncbi.nlm.nih.gov/pubmed/22155335

[7] http://wordnet.princeton.edu/

2. Evaluations and conclusion

As a test case, we align EMMeT[8] to Dorland's (32nd edition). We evaluate a random sample of non-exact alignments (100), comparing them with a baseline Jaro-Winkler (JW) matching approach. AnAGram gives more correct results and JW finds more related matches (Table 1- top two lines, and Figure 1).

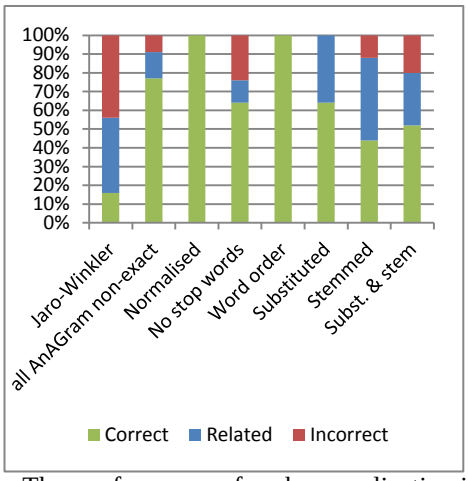

Figure 1 - Quality of matches returned by AnAGram's modules.

Preferred labels	C	R	I
Jaro-Winkler	16	40	44
AnAGram non-exact	77	14	9
Normalised	25	0	0
No stop words	16	3	6
Word order	25	0	0
Substituted	16	9	0
Stemmed	11	11	3
Subst. & stem	13	7	5

Table 1 – Results for AnAGram's modules. (C: correct; R: related; I: incorrect)

The performance of each normalization is evaluated using 25 random results for each of AnAGram's modules separately[9] (Table 1- bottom, Figure 1). Normalization does very well (100% correct results). Removal of stop words causes some errors and related matches (stop words can be meaningful like *A* for *hepatitis A*). Word order rearranging ranks second: it does not often change the meaning of the term. Substitution performs reasonably well: most of the non-correct results are related matches. Stemming gives the poorest results, with false positives due to nouns/verbs stemmed to the same root, such as *cilitated/ciliate*. The substituted-and-stemmed matches have a result similar to the stemmed results. Still, even the worst results from any AnAGram module are better than the overall results of the non-exact matches from the JW algorithm. One reason for this can be that JW does not stop the alignment at the best match, but delivers everything that satisfies the threshold.

Not all modules account for an equal portion of the non-exact results. The normalization module delivers around 70% of matches, stemming accounts for 15 to 20% and the other modules account for 2% to 4% of the matches each.

AnAGram's results are good compared to the performance of string-based methods in the OAEI large biomedical vocabularies alignment[10]. We will work on the Stemming algorithm, on improving our stop words list and substitution dictionary, and on adding an optimized version of the JW algorithm, thus benefitting from additional related matches where no previous match was found.

[8] Version 3.2, from December 2013

[9] Some modules use previous transformation results.

[10] http://oaei.ontologymatching.org/2013/largebio/index.html