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: Biomedical vocabularies have specific characteristics that make their lexical alignment challenging. We have built a string-based vocabulary alignment tool, AnAGram, dedicated to efficiently compare terms in the biomedical domain, and evaluate this tool's results against an algorithm based on Jaro-Winkler's edit-distance. AnAGram is modular, enabling us to evaluate the precision and recall of different normalization procedures. Globally, our normalization and replacement strategy improves the F-measure score from the edit-distance experiment by more than 100%. Most of this increase can be explained by targeted transformations of the strings with the use of a dictionary of adjective/noun correspondences yielding useful results. However, we found that the classic Porter stemming algorithm needs to be adapted to the biomedical domain to give good quality results in this area.

1. Introduction

Elsevier has a number of online tools in the biomedical domain. Improving their interoperability involves aligning the vocabularies these tools are built on. The vocabulary alignment tool needs to be generic enough to work with any of our vocabularies, but each alignment requires specific conditions to be optimal, due to vocabularies' specific lexical idiosyncrasies.

We have designed a modular, step-wise alignment tool: AnAGram. Its normalization procedures are based on previous research[1], basic Information Retrieval normalization processes, and our own observations. We chose a string-based alignment method as these perform well on the anatomical datasets of the OAEI campaign[1], and string-based alignment is an important step in most methods identified in [3][4].

We compare the precision and recall of AnAGram against an implementation of Jaro-Winkler's edit-distance method (JW)[7] and evaluate the precision of each step of the alignment process. We gain over 100% F-measure compared to the edit-distance method. We evaluate the contribution and quality of the string normalization modules independently and show that the Porter stemmer[2] does not give optimal results in the biomedical domain.

In Section 2 we present our use-case: aligning Dorland's to Elsevier's Merged Medical Taxonomy (EMMeT)[1]. Section 3 describes related work in vocabulary

[1] http://river-valley.tv/elsevier-merged-medical-taxonomy-emmet-from-smart-content-to-smart-collection/

alignment in the biomedical domain. Section 4 and 5 present AnAGram and evaluate against Jaro-Winkler's edit-distance. Section 6 presents future work and conclusions.

2. Use case: Dorland's definition alignment to EMMeT

Elsevier's Merged Medical Taxonomy (EMMeT) is used in "Smart Content" applications[2]; it contains more than 1 million biomedical concepts and their hierarchical, linguistic and semantic relationships. We aim at expanding EMMeT with definitions from the authoritative biomedical dictionary Dorland's[3] by aligning them.

3. Related work

Cheatham and Hitzler[1] list the types of linguistic processes used by at least one alignment tool in the Ontology Alignment Evaluation Initiative (OAEI)[5]. AnAGram implements all syntactic linguistic transformations listed; instead of a generic synonym expansion system, we used a correspondence dictionary of adjective/noun pairs. This dictionary is a manually curated list based on information automatically extracted from Dorland's. It contains pairs that would not be not solved by stemming such as *saturnine/lead*. Ambiguous entries, such as *gluteal/natal,* were removed.

Chua and Kim's[6] approach for string-based vocabulary alignment is the closest to AnAGram: they use WordNet[4], a lexical knowledge base, to gather adjective/noun pairs to improve the coverage of their matches, after using string normalization steps; our set of pairs is larger than the one derived from WordNet.

4. AnAGram: biomedical vocabularies alignment tool

AnAGram was built for use on a local system[5], and is tuned to performance by using hash-table lookup to find matches. Currently, no partial matching is possible. The matching steps are built in a modular way: one can select the set of desired steps. The source taxonomy is processed using these steps and the target taxonomy is processed sequentially: the alignment stops at the first match. Modules are ordered to increasing distance between original and transformed string, simulating a confidence value.

Exact matching: corresponds to JW edit-distance 1.

Normalization: special characters are removed or transformed (*Sjögren's syndrome* to *Sjogren's syndrome*; punctuation marks to space), string is lower cased.

Stop word removal: tokenization by splitting on spaces, removal of stop words, using a list that was fine-tuned over several rounds of indexing with EMMeT.

[2] http://info.clinicalkey.com/docs/Smart_Content.pdf

[3] http://www.dorlands.com/

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

[5] 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

Re-ordering: tokens are sorted alphabetically, enabling matches for inverted terms.
Substitution: sequences of tokens are replaced with the corresponding value from our dictionary, applying a longest string matching principle.
Stemming: using the Porter stemming algorithm[2] (Perl module Lingua::Stem:: Snowball). The substitution step is then repeated, using stemmed dictionary entries.
Independent lists: stop-words list and substitution dictionary are independent files.

5. Experimentation and results

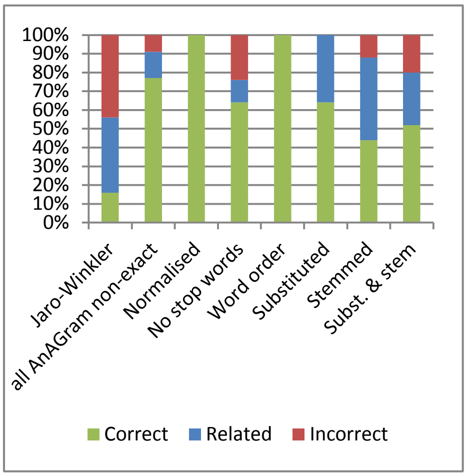
We align EMMeT version 3.2 (13/12/13) (1,027,717 preferred labels) to Dorland's 32nd edition (115,248 entries). We evaluate AnAGram as a whole against JW, with a 0.92 threshold (established experimentally). The JW implementation can work only with preferred labels.

To evaluate the recall of AnAGram vs the JW implementation, we use a manual gold set of 115 mappings created by domain experts (Table 1). AnAGram gives better recall and better precision than the JW method.

	Correct mapping	Incorrect mapping	Recall (%)	Precision(%)	F-measure
Jaro-Winkler	46	8	43%	85%	0.57
AnAGram	80	3	71%	96%	0.82

Table 1 - Results of AnAGram vs. Jaro-Winkler on Dorland's Gold Set pairs

We evaluate a random sample of 25 non-exact alignments from each module to get a better insight on AnAGram's normalization process. The results are either: *Correct*, *Related* (useful but not exactly correct), or *Incorrect* (Table 2 and Figure 1). AnAGram gives more correct results but JW is useful for finding related matches.



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

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

We evaluate the performance of each normalization step by evaluating 25 random results for each of AnAGram's modules separately[6] (Table 2, Figure 1). Normal-

[6] Some modules are based on the result of a previous transformation, so the later the module comes in the chain, the more complicated matches it faces.

ization does very well (100% correct results). Removal of stop words causes some errors and related matches: single-letter 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 is that the JW does not stop the alignment at the best match, but delivers everything that satisfies the threshold of 0.92.

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.

6. Future work and conclusion

Results are good compared to OAEI large biomedical vocabularies alignment's results for string-based tools[1]. We will work on the Stemming algorithm, the improvement of our stop words list and substitution dictionary, and on adding an optimized version of the JW algorithm as a final optional module for AnAGram to improve results further. In this way we will benefit from additional related matches in cases where no previous match was found.

References

[1] Michelle Cheatham, Pascal Hitzler. *String Similarity Metrics for Ontology Alignment*. International Semantic Web Conference (ISWC2013) (2) 2013: 294-309

[2] Cornelis .J. van Rijsbergen, Stephen E. Robertson, MartinF. Porter. *New models in probabilistic information retrieval.* London: British Library. (British Library Research and Development Report, no. 5587), 1980

[3] Jérôme Euzenat (Coordinator) et al. *State of the art on Ontology alignment*. Knowledge Web D 2.2.3, 2004.

[4] Jerôme Euzenat, Pavel Shvaiko. *Ontology Matching*. Springer-Verlag, Berlin Heidelberg 2013

[5] Jérôme Euzenat, Christian Meilicke, Heiner Stuckenschmidt, Pavel Shvaiko, Cássia Trojahn. *Ontology Alignment Evaluation Initiative: Six Years of Experience*. Journal on Data Semantics XV, Lecture Notes in Computer Science (6720) 2011: 158-192

[6] Watson W. K. Chua and Jung-Jae Kim. *BOAT: Automatic alignment of biomedical ontologies using term informativeness and candidate selection*. Journal of Biomedical Informatics (45) 2012: 337-349

[7] William E.Winkler. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. Proceedings of the Section on Survey Research Methods (American Statistical Association) 1990: 354–359