Mining and Ranking Biomedical Synonym Candidates from Wikipedia

Abhyuday N Jagannatha
College of Information and Computer Sciences,
University of Massachusetts, Amherst
MA 01003, USA
abhyuday@cs.umass.edu

Jinying Chen
Department of Quantitative Health Sciences,
University of Massachusetts, Worcester
MA 01605, USA
jinying.chen@umass-med.edu

Hong Yu
Veterans Administrative Medical Center, Bedford
MA 01730, USA
hong.yu@umassmed.edu

Abstract

Biomedical synonyms are important resources for Natural Language Processing in Biomedical domain. Existing synonym resources (e.g., the UMLS) are not complete. Manual efforts for expanding and enriching these resources are prohibitively expensive. We therefore develop and evaluate approaches for automated synonym extraction from Wikipedia. Using the inter-wiki links, we extracted the candidate synonyms (anchor-text e.g., "increased thirst") in a Wikipedia page and the title (e.g., "polyuria") of its corresponding linked page. We rank synonym candidates with word embedding and pseudo-relevance feedback (PRF). Our results show that PRF-based reranking outperformed word embedding based approach and a strong baseline using inter-wiki link frequency. A hybrid method, Rank Score Combination, achieved the best results. Our analysis also suggests that medical synonyms mined from Wikipedia can increase the coverage of existing synonym resources such as UMLS.

1 Introduction

Biomedical synonym resources have been an important part of biomedical natural language processing (NLP). Synonym resources have been used for a variety of tasks such as query expansion (Aronson and Rindflesch, 1997; Díaz-Galiano et al., 2009), reformulation (Plovnick and Zeng, 2004), and word sense disambiguation (McInnes et al., 2007).

Another important avenue of their use lies in e-portals for clinical notes such as My HealtheVet patient portal, which allows patients to access clinical notes written by their healthcare providers (Nazi et. al., 2013). While many organizations have been embracing these methods of patient-clinician communication, various studies (Lerner et al., 2000; Chapman et al., 2003; Keselman et al., 2007) have shown that patients often have difficulty in comprehending clinical notes.

A patient's ability to comprehend clinical notes is directly related to his/her ability to understand medical jargon (Pyper et al., 2004; Keselman et al., 2007). Subsequently approaches have been developed to replace medical jargon with corresponding lay terms (Kandula et al., 2010; Abrahamsson et al., 2014). Such approaches rely on high quality synonym resource(s). The widely used biomedical knowledge resource, Unified Medical Language System (UMLS) (Humphrey et al., 1998) is a very valuable resource for such purposes. The UMLS incorporates over 100 biomedical terminology resources including Consumer Health Vocabulary (CHV). It also contains definitions for medical terms which can be used to simplify the clinical notes (Ramesh et al., 2013). Even though UMLS is a rich resource with a vast quantity of medical terms, we found that several synonymous or related medical terms that we extracted through Wikipedia, were not present in the UMLS dictionaries. We report this coverage in Section 5.2.

In this paper, we propose a data-driven approach for automatic extraction and ranking of medical synonyms from Wikipedia. Wikipedia is a free-access, free-content collaborative online encyclopedia. Our previous work suggests that about 40% content in Wikipedia contain health related information (Liu et al., 2013). Many studies have shown that Wikipedia contains high quality of biomedical content (Reavley et al., 2012; Devgan et al., 2007; Rajagopalan et al., 2011). For

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 142–151,
Lisbon, Portugal, 17 September 2015.

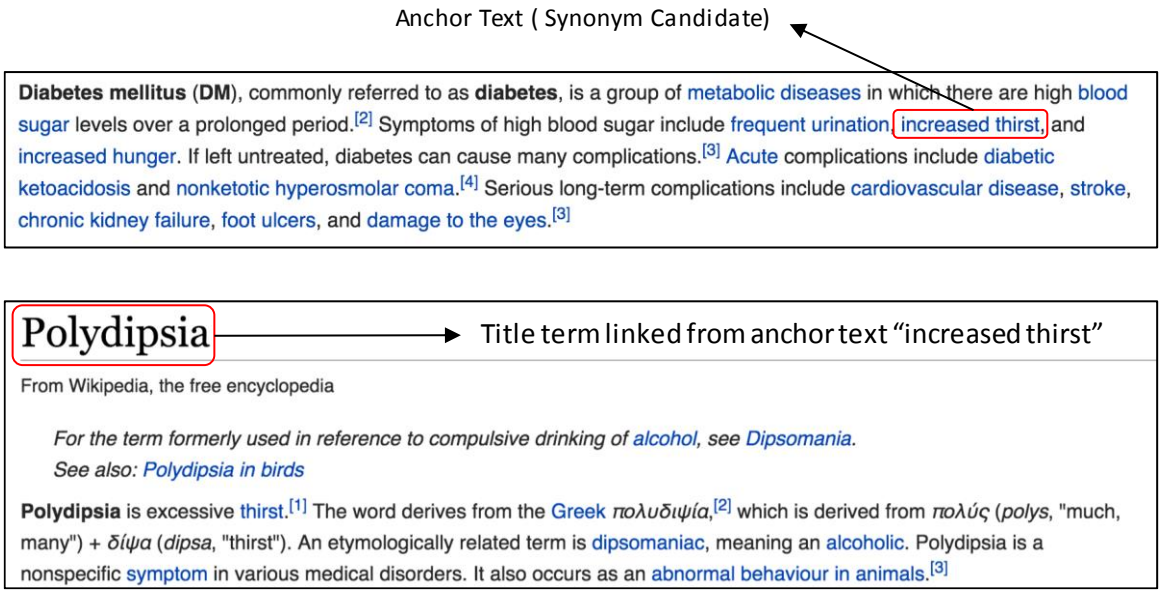

Figure 1: Introductory paragraph for the "Diabetes Mellitus" page in English Wikipedia on top, along with the "Polydipsia" Page below it

example, Devgan et al. (2007) evaluated that Wikipedia contains highly accurate medical articles. They do however, also mention that some articles contain incomplete medical content. Rajagopalan et al. (2011) concluded that Wikipedia has similar accuracy and depth as a professionally edited database. Similarly, Reavley et al. (2012) showed that Wikipedia contains high quality information on mental disorders.

As the result, Wikipedia is being increasingly used by healthcare providers. Specifically, studies show that Wikipedia is widely used by junior physicians (Hughes et al., 2009) and pharmacists (Brokowski & Sheehan, 2009). Additionally Wikipedia is also being widely used by people who are looking for healthcare information. Based on the search engine ranking and page view statistics, Laurent and Vickers (2009) concluded that English Wikipedia is major source of health related information for online users.

Since Wikipedia is written collaboratively by anonymous volunteers, a majority of whom are lay people, its content contains both biomedical jargons and lay terms. This makes Wikipedia a rich resource linking medical jargon with synonymous lay phrases. We leverage this resource by extracting inter-wiki links from Wikipedia to obtain (page title, anchor text) pairs. A typical Wikipedia page includes a title and a description text in which anchor texts are linked (through inter-wiki links) to other Wikipedia pages. As illustrated in Figure 1, one of the anchor texts in the "Diabetes mellitus" Wikipedia page, "increased thirst" is linked to the corresponding page with the title term "polydipsia". We treat the anchor text as a synonym candidate for the title term, which we treat as target concept.

Synonym candidates and their target concepts extracted from inter-wiki links are often synonymous pairs. For example, the anchor texts "frequent urination," "increased thirst," and "increased hunger" are linked to the title pages of "polyuria," "polydipsia," and "polyphagia", respectively. However, sometimes, the synonym candidates and their target concepts are only related but not synonymous. For example "nonketotic hyperosmolar coma" and "kidney failure" are linked to the "hyperosmolar hyperglycemic state" and "chronic kidney disease" respectively.

In addition, as a crowdsourcing resource, Wikipedia has noise. A typical case is where the inter-wiki links are tagged partially. For example, only the "attack" in "heart attack" may be linked to "Myocardial Infarction".

To improve the quality of synonym extraction, we explored several unsupervised methods to rank the synonymous pairs, which utilize distributed word representation (i.e., word embeddings), pseudo relevance feedback (PRF) based re-ranking, and ranking combinations. To our knowledge, this is the first effort that uses word embedding-based ranking and PRF to improve

synonym extraction from Wikipedia. We compared our methods with a strong baseline method which uses entity-link frequency.

2 Background

2.1 Related Work

Synonym identification has been active research for two decades. Landauer and Dutnais (1997) used latent semantic analysis to generate 300-dimension word vectors to rank answers of synonym questions in TOEFL. Turney (2001) used search queries to obtain Point-Wise Mutual Information score for two terms to judge whether they are synonyms. Yu et al. (Yu et al., 2002; Yu and Agichtein, 2003) developed rule-based and learning-based methods for extracting author-defined synonyms from text (e.g., using surface cue phrases such as "also called" and parentheses to identify full synonyms and their abbreviations).

Neelakantan and Collins (2015) applied Canonical Correlation Analysis to calculate representation of phrases which were then used for synonym classification. McCrae and Collier (2008) used automatically generated patterns (regular expressions) to mine candidate synonym pairs, which were then classified as synonymous or not based on the occurrence of term pairs in each pattern. Henriksson et al. (2014) created ensembles of semantic spaces, by combining different distributional models and semantic spaces induced from different corpora, for synonym extraction. Blondel et al. (2004) used a central similarity measure in word graphs to calculate similarity between two words. They constructed their graph using a dictionary with the assumption that synonyms were likely to have common words in their definitions and might simultaneously appear in the definitions of many other words. Wang et al. (2015) modified the word2vec algorithm to create a semi-supervised approach that learned from both unlabeled text corpus and UMLS semantic types, groups.

Bøhn and Nørvåg (2010) used redirect pages and inter-wiki links to extract named entities from Wikipedia. They used the frequency of inter-wiki links and other heuristics (e.g., letter capitalization) to rank the synonym candidates. In our work, we use inter-wiki link frequency as our baseline and study the improvements provided by various methods described in section 3.

2.2 Word Representations

Word representations keep the semantic and contextual information of a word in a compact format (e.g., a vector or a tensor). Different methods have been used to obtain compact representations, including clustering based approaches (e.g., Brown Clustering (Brown et al., 1992)), co-occurrence based approaches (Lebret and Collobert, 2014; Pennington et al., 2014), and hierarchical language models (Mnih and Hinton, 2009). Mikolov et al. (2013a, 2013b) showed that using a dense vector representation for words outperforms methods like tf-idf in NLP tasks, e.g., Microsoft Sentence Completion Challenge (Zweig and Burges, 2011). It is expected that words sharing similar semantics or contexts will be close in the projected latent space. In this study, we used the Skip Gram model (Mikolov et al. 2013a) to compute relatedness of synonym pairs extracted from the Wikipedia. Skip Gram models, which belong to distributed word representation (i.e., word embedding) models, are trained through a log-linear classifier that maximizes the prediction accuracy of words within a certain range before and after the current word. We used word vector based similarity methods to rank the synonym candidates because we believe that it has a better semantic representation than the simpler frequency-based approach.

Medical target concepts in Wikipedia are often linked to a variety of synonym candidates; however we found that for several cases, the number of links for each synonym candidate sometimes is very low. For those cases, frequency of inter-wiki links may not be sufficient to accurately determine the ranking of synonym candidates. For example, the target concept "myopathy", is linked to "exertional myopathy", "hereditary myopathy", "muscle disorders", "muscle weakness", "muscular diseases", "polyneuropathy" and "metabolic myopathy", "progressive myopathy" through inter-wiki links. However, each of these inter-wiki links occurs only once. As a consequence, we cannot rank these synonym candidates using their link frequencies. Word embedding approaches do not suffer from this problem and are expected to perform better in such cases.

In addition, word embedding approaches can filter out frequent but partial synonym candidates and provide better ranking. An example of a partial synonym candidate is the "heart attack" example discussed before, where only the word "attack" is tagged as the anchor-text. We expect that such erroneous synonym candidates are rare occurrences and can be filtered out using their link frequency. But, in reality due to erroneous manual tagging, partial anchor-texts (i.e. synonym candidates) sometimes occur more frequently than the

true synonyms. For example, "oral cancer" is linked most frequently through "mouth" and "oral" (eight and four times respectively), while a correct paraphrase like "cancer of mouth and tongue" is only linked one time. Word embedding approaches represent semantics better than the frequency-based approach and therefore may be able to identify synonyms and separate them from false positives.

2.3 Pseudo Relevance Feedback

We use pseudo-relevance feedback (PRF) (Attar and Fraenkel, 1977), a widely used method in information retrieval (IR), to obtain better estimates of the representations of target concept in the latent space. PRF is a subtype of a broader class of methods called relevance feedback models (Rocchio, 1971) in IR. Relevance feedback models exploit the idea of using feedbacks (typically from the user) about the relevancy of the results returned for an initial query, to improve or enrich this query. PRF, in particular, does not require user interaction, but instead uses the top-*k* retrieved documents as an automatic feedback. These top-ranked documents are added to the query, and the search runs again with the updated query. We adapted this approach to solve the problem of ranking synonym candidates, which we will introduce in detail in Section 3.3

3 Methods

To improve synonym extraction from Wikipedia inter-wiki links, we explored different unsupervised approaches, including several new methods, for synonym candidate ranking.

3.1 Entity Link Frequency (ELF)

ELF ranks (target concept, synonym candidate) pairs by their Wikipedia inter-wiki link frequency. More specifically, each synonym candidate is ranked by the number of times it has been used as an anchor-text to link to the target concept. Because the inter-wiki links in Wikipedia are created manually, the link frequency associated with each candidate term is a very strong indicator of the viability of that particular synonym candidate. Noisy inter-wiki links (e.g., "arrhythmia" — "other causes" and "heart attack" — "attack") often have low frequencies; while high frequency terms ("polydipsia" — "excessive thirst") are often good synonym candidates. This method was used as the baseline in our experiments.

3.2 Word-Embedding Based Ranking

We use word vectors to estimate the similarity of two words by computing the cosine similarity of their vectors in the embedded space. Many medical terms, however, are phrases with two to five words. This requires methods to combine individual word vectors into phrases. In this work, given two phrases *a* and *b* (represented by "a_1 ... a_n" and "b_1 ... b_m" respectively), we estimate their similarity by using the average cosine distance between each pair of words they contain, as defined in Equation 1,

$$ACS_{ab} = \frac{1}{nm}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} < W(a_i), W(b_j) >\right) \quad (1)$$

where $W(.)$ is the normalized word vector of an individual word. This can be interpreted as computing the cosine similarity of the two phrase vectors, where a phrase vector is estimated by the mean of the normalized word vectors of the individual words contained in that phrase. We call this method Average Cosine Similarity (ACS).

3.3 Re-ranking based on Pseudo Relevance Feedback (PRF)

A limitation of the word embedding method that we use, Skip Gram model, is that it does not disambiguate word senses. In other words, the vector of a word represents multiple senses of this word. As a consequence, synonym candidates with non-relevant senses (e.g., a non-medical sense of the target concept) could be ranked high by word embedding-based ranking method. To alleviate this problem, we leverage on Relevance feedback to disambiguate our term vectors.

As introduced in the previous section, Pseudo Relevance Feedback (Attar and Fraenkel, 1977), is a popular technique in IR, which expands a given query by the top-n documents retrieved for this query. This updated query is then used to retrieve the documents. We adapted PRF for our problem by collecting the top-n synonym candidates obtained by the ELF method. We then calculate the mean vector of these *n* candidate phrases and the target concept. This mean vector is used as the new query. We then re-rank all synonym candidates by their Average Cosine Similarity (ACS) to this new query. When selecting the top-*n* synonym candidates through ELF, if there are multiple candidates with the same ELF scores, we use ACS to break the tie. For example, if the

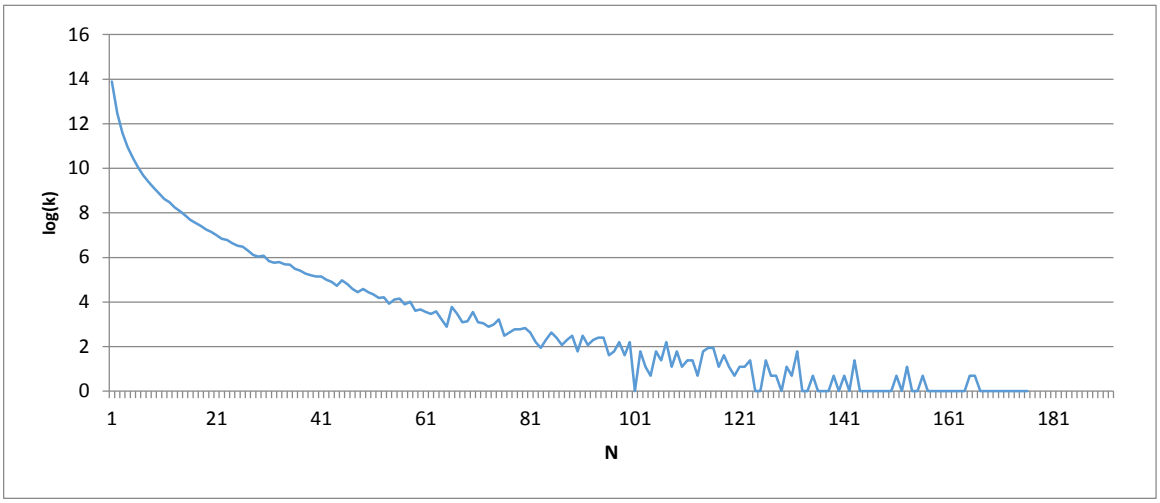

Figure 2: Distribution of number of synonym candidates for Wikipedia Terms. *N* is the number of synonym candidates extracted per title term. Wikipedia Concept. *k* is the number of title terms in Wikipedia that have *N* unique number of synonym candidates

synonym candidates "blood infection" and "bacterial infection" for the target concept "septicemia" have ELF score 1, the candidate with the higher ACS to the concept term will be chosen with a higher priority. In case ACS scores are also tied, we randomly order them.

One major advantage of this method is the use of mean vector of the top-*n* candidates to represent the dominant sense of the target concept. Therefore, it will rank synonym candidates that have senses similar to this dominant sense very highly.

3.4 Ranking Combination

Ensemble ranking is a standard method to combine the strengths of different ranking methods. When we conducted this work, we did not have any annotated data to build a standard ensemble ranker. Instead, we adopted two simple unsupervised methods for ranking combination: (1) Average Ranking (AveR); and (2) Ranking Score Combination (RSC).

AveR ranks the candidates by their mean ranks from ELF, ACS and PRF. RSC ranks the candidates by the sum of their ranking scores from ELF, ACS and PRF. We did not normalize the ELF scores into [0,1] by observing that a large ELF score (i.e., high inter-wiki link frequency) often correctly predicts synonyms, irrespective of its corresponding ACS or PRF scores (which are values that lie between 0 and 1). Our preliminary experiments comparing score combination using normalized vs. raw ELF scores confirmed our choice of using raw ELF scores.

4 Experimental Settings

4.1 Experimental Data

We extracted all the (target concept, synonym candidate) pairs from Wikipedia except the pairs that contain special characters or numbers. In total, we obtained 24M links, with 3.6M unique links for 1.6M distinct concepts.

Figure 2 shows the distribution of the number of synonym candidates extracted for each title term from the Wikipedia. Out of the total 1,659,049 title-terms, 1,457,935 terms have less than three synonym candidates. Our preliminary study suggests that many of these terms are person or location names, which are not of our interest. Therefore, we did not include these terms when creating our gold-standard evaluation dataset and only evaluated our methods on terms with three or more synonym candidates.

We use word2vec software to create the Skip Gram word embeddings. The word embeddings were trained on a combined text corpus of English Wikipedia, Simple English Wikipedia and articles from PubMed Open Access, which contain over 4 billion words in total. The text was lowercased and stripped of all punctuations except comma, apostrophe and period.

We set our word2vec training parameters based on the study of Pyysalo et al. (2013). Specifically, we used 200-dimension vectors with a window size of 6. We used hierarchical soft-max with a subsampling threshold of 0.001 for training.

4.2 Evaluation Dataset

There is no lexical resource suitable for evaluating our task performance. Even UMLS does not cover all the synonyms and related terms we discovered from the Wikipedia. To evaluate our synonym ranking methods, we created a gold standard evaluation dataset from the Wikipedia data we extracted.

Since the goal of this work is to extract synonym candidates for medical terms, we only chose medically relevant concepts for evaluation. We randomly selected 4000 terms from the concepts (title terms) that are present either in the Consumer Health Vocabulary or in the Wikipedia Health Category tree to the depth of 4. An annotator with PhD degree in Biology further selected 1000 relevant medical terms from these 4000 terms.

We built an annotation GUI that presented to the annotators 1000 medical terms and their synonym candidates. Each term and its synonym candidates were shown in a single annotation page. The page order was randomized. The annotation task was to judge whether the synonym candidate was a "Synonym", "Related Term" or "Rejected or Unrelated Term" of the target concept. Two annotators conducted the annotation. Both are premedical school students. So far, 792 unique medical concepts were annotated, out of which 256 were annotated by both. We used these 256 concepts for our evaluation. We also used the entire 792 concepts and their synonyms to calculate the coverage by UMLS.

A synonym candidate is defined as a "Synonym" if it has the exact same meaning as the target concept. It is defined as a "Related Term" if it has a related meaning to the target concept. We accept hypernyms, hyponyms, and words derived from the same root as "Related terms". Additionally we also accept words with high correlations to the target concept, e.g., a very common symptom for a disease. As an example, "high blood sugar" is a related term of "diabetes mellitus". Candidates not in the above-mentioned categories were annotated as "Unrelated or Rejected Terms".

The gold standard of 256 concepts consists of 1507 (title term, synonym candidate) pairs and their corresponding annotations. The linear weighted kappa for the inter-annotator agreement was 0.4762, with the 95% confidence interval ranges from 0.4413 to 0.5111. This kappa value suggests that the annotators have moderate agreement (Viera and Garret, 2005). If we combine related terms and rejected terms into one category, the resulting dataset has a much higher kappa of 0.6250. This contrasts with a low kappa of 0.3929 when related terms are instead combined with synonyms, suggesting that more annotator uncertainty lies in the boundary between related and rejected terms than between related and synonymous terms.

4.3 Evaluation Measure

We use mean average precision (MAP) to evaluate the performances of our ranking methods, because our problem is similar to a typical Information Retrieval tasks. Instead of using a set of relevant and irrelevant documents to evaluate our ranking output, we use a set of synonyms, related terms and rejected terms from our gold-standard annotation for evaluation.

We set two evaluation conditions: (1) combining the synonyms and related terms from the gold-standard annotation to form the set of relevant (positive) instances and treating rejected terms as irrelevant (negative) instances; and (2) using the synonyms from the gold annotation as positive (relevant) instances and treating the related and rejected terms as irrelevant (negative) instances.

By the above definition, condition 1 is a relaxed condition and condition 2 is strict. For both conditions, only terms that were judged by both annotators as relevant (positive) instances are treated as positive.

We compute MAP by Equations (2) and (3).

$$AveP = \sum_{k=1}^{n} P(k)\Delta_r(k) \quad (2)$$

$$MAP = \frac{\sum_{t=1}^{M} AveP(t)}{M} \quad (3)$$

where *AveP* is the average precision of a query (target concept in our case); *k* is the rank of the synonym candidates; *P(k)* is the precision of the ranking at rank *k*; $\Delta_r(k)$ is the increase of recall of the ranking at rank *k* compared with the recall at rank *k-1*; *MAP* is the mean *AveP* of all the target concepts to evaluate on.

5 Results

5.1 Synonym Candidate Ranking

The ranking performances (measured by MAP) of different methods are shown in Table 1.

As we see, under the relaxed condition (Column 1), the word embedding-based ranking method ACS outperforms the frequency based ranking method ELF (Row 2 vs. Row 1); while under the strict condition (Column 2), ACS has slightly lower performance than ELF (Row 2 vs.

Methods	*MAP (Relaxed condition)*	*MAP (strict condition)*
ELF	0.6267	0.2401
ACS	0.6624	0.2383
PRF	0.6859	0.2519
AveR	0.6685	0.2433
RSC	**0.6900**	**0.2745**

Table 1: Mean Average Precision values for Relevance Feedback of 5

Row 1). This suggests that the word embedding-based ranking method is superior than the frequency based ranking method in identifying semantically related (coherent) terms. However, they themselves may not be sufficient to accurately identify synonyms.

Result analysis suggests that ELF has a high precision at high ranks, especially when the frequency of the candidate term (i.e., the number of times it is linked to the target term in Wikipedia) is high. However, the frequency values for synonym candidates tend to be identical for lower ranked candidates. As a result, it is impossible to determine the order of these candidates using frequency based method such as ELF. Table 2 shows a typical example. For the target concept "septicemia", the frequencies of its candidates are all 1's. In this case, we cannot gain any information from ELF about the ranking of these synonym candidates. The annotated rankings from one of our annotators and the rankings predicted by the PRF method are given on the side. This is a major reason why ELF has lower MAP than ACS and PRF.

PRF performs better than ELF and ACS consistently on both conditions. As introduced in Section 3, in the PRF method, we use the top-*n* (*n*=5 in our experiments) candidate terms returned by the ELF method as feedback terms and use ACS to break the tie (when there are candidates with the same ELF scores). This way, PRF implicitly takes advantages of both ELF and ACS, which explains why it is better than these two methods.

Further analysis of the results suggest that PRF is good at rejecting unrelated terms, but can be confused between synonyms and related words. This is especially true when the related terms are just morphologically different from the original term (see Table 2 for an example).

Table 1 also shows the performance from combining individual ranking methods. As we can see, the performance of the average ranking method using equal weights (AveR, Row 4) falls between the best and the worst individual ranking methods on both conditions. This is not surprising because

Candidates for "septicemia"	*Annotation*	*PRF Ranking*	*Frequency*
bacterial infection	Related	2	1
blood infection	Synonym	6	1
coral poisoning	Rejected	7	1
Septicaemia	Synonym	1	1
Septicaemic	Related	4	1
Septicemic	Related	3	1
septic infection	Synonym	5	1

Table 2: Predictions for "septicemia"

we did not tune the combination weights. It is likely we can boost the ranking performance by optimizing the combination weights using annotated training data, which will be our future work. Interestingly, the performance from using combined ranking scores (RSC) is almost always higher than all the individual methods with respectable margins on both conditions. This result suggests that augmenting ELF rankings with word similarity based measures and pseudo relevance feedback is a very effective way to improve the quality of synonym candidate ranking. Paired t-test shows that our best performing method RSC is significantly better than the ELF baseline on both conditions (p- value<0.001). Other methods are significantly better than ELF on the relaxed condition (p-value <0.01) but not on the strict condition.

In our experiments, we set *n*, the number of feedback terms used by PRF, as 5. This value was set heuristically due to the lack of the training data. In a post-experiment analysis, we tested the effects of using different values of *n* (from 1 to 10). Figures 3 and 4 show the results. As we can see, the values of *n* do not affect the ranking results remarkably, especially on the strict condition. In particular, the orders of the performances of different methods remain the same.

5.2 Coverage of Synonym Extraction

To estimate how much the Wikipedia based synonym extraction can contribute to existing synonym resources, we analyzed the coverage of our synonyms in UMLS. So far, we have 5025 unique pairs of medical concepts and their synonym terms, which have been annotated (judged) by at least one annotator. Of the 5025 pairs, 4447 have been annotated as either a synonym or a related term. Of these 4447 terms, only 2621 are covered in UMLS.

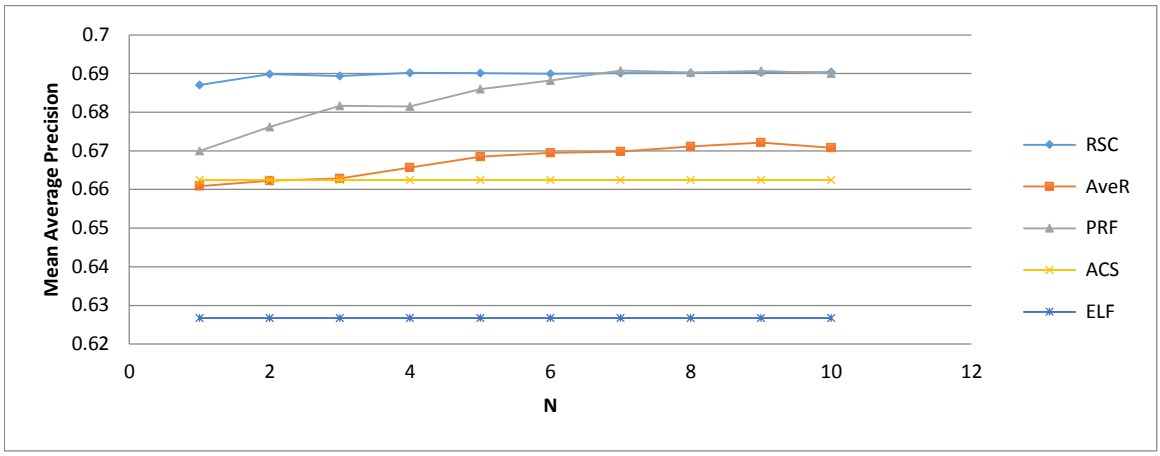

Figure 3: Plot of Mean Average Precision vs N for relaxed condition. N is the number of queries used for Relevance Feedback

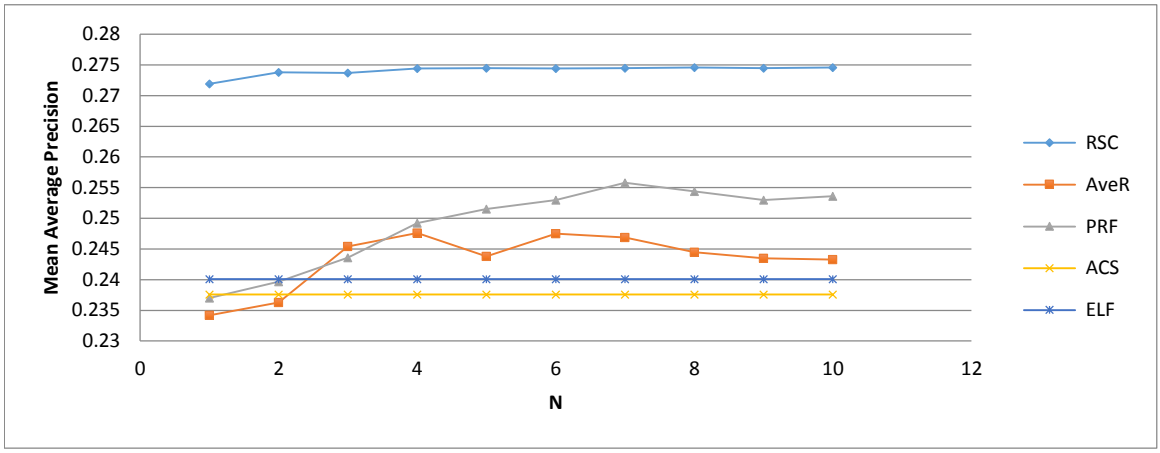

Figure 4: Plot of Mean Average Precision vs N for strict condition. N is the number of queries used for the Relevance Feedback

If we look at only synonyms, 1523 of the 5025 pairs have been annotated by at least one of the annotators as synonyms. Out of the 1523 terms, 429 are not in UMLS.

Clearly Wikipedia is a valuable synonym resource that can be mined to enhance existing lexical resources such as UMLS.

6 Conclusion and Future Work

We have presented novel methods in mining and ranking synonyms from Wikipedia. Our approach is distinguished from previous works in that we utilize word embeddings and pseudo relevance feedback to estimate the semantic and contextual similarities of medical terms and use them as a feature to improve synonym candidate ranking. Our results show that a combination of frequency-based ranking, word embedding based ranking and pseudo relevance feedback achieves the best performance. This suggests that word embedding is a valuable tool in improving synonym extraction from noisy resources like Wikipedia.

We used English Wikipedia for this work. Our approach is general and can be applied to other languages. Its performance is contingent on the size of the Wikipedia and the quality of word embeddings for each specific language. Wikipedia has more than 280 languages, 50 of which have more than hundreds of thousands of articles. The word2vec tool can be trained and used on corpora in any of these languages.

We use the mean of individual word vectors to estimate the phrase vector. In the future, we will explore more advanced algorithms (e.g., Recursive Neural Networks (Socher et al., 2011)) for phrase composition.

The synonym pairs mined and ranked by our methods will be added to a comprehensive synonym resource after manual curation. We will use this resource to simplify medical health records, by substituting complex medical terms with their lay language synonyms.

Acknowledgements

We would like to thank our annotation team (Elaine Freund, Victoria Wang and Shreya Makkapati) for creating the gold-standard evaluation set used in this work.

This work was supported in part by the Award 1I01HX001457 from the United States Department of Veterans Affairs Health Services Research and Development Program Investigator Initiated Research. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

References

Abrahamsson, E., Forni, T., Skeppstedt, M., & Kvist, M. (2014). Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 57–65.

Aronson, A. R., & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. *Proc AMIA Annu Fall Symp*, 485–9.

Attar, R., & Fraenkel, A. S. (1977). Local Feedback in Full-Text Retrieval Systems. *J. ACM*, *24*(3), 397–417.

Blondel, V., Gajardo, A., Heymans, M., Senellart, P., & Van Dooren, P. (2004). A measure of similarity between graph vertices. *arXiv:cs/0407061*.

Bohn, C., & Nørvåg, K. (2010). Extracting Named Entities and Synonyms from Wikipedia. In *Advanced Information Networking and Applications* (pp. 1300–1307).

Brokowski, L., & Sheehan, A. H. (2009). Evaluation of pharmacist use and perception of Wikipedia as a drug information resource. *The Annals of Pharmacotherapy*, *43*(11), 1912–1913.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–479.

Chapman, K., Abraham, C., Jenkins, V., & Fallowfield, L. (2003). Lay understanding of terms used in cancer consultations. *Psycho-Oncology*, *12*(6), 557–566.

Devgan, L., Powe, N., Blakey, B., & Makary, M. (2007). Wiki-Surgery? Internal validity of Wikipedia as a medical and surgical reference. *Journal of the American College of Surgeons*, *205*(3, Supplement), S76–S77.

Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2009). Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in Biology and Medicine*, *39*(4), 396–403.

Henriksson, A., Moen, H., Skeppstedt, M., Daudaravičius, V., & Duneld, M. (2014). Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, *5*(1), 6.

Hughes, B., Joshi, I., Lemonde, H., & Wareham, J. (2009). Junior physician's use of Web 2.0 for information seeking and medical education: A qualitative study. *International Journal of Medical Informatics*, *78*(10), 645–655.

Humphrey, B., Lindberg, D. A. B., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Association*, *5*, 1–11.

Kandula, S., Curtis, D., & Zeng-Treitler, Q. (2010). A Semantic and Syntactic Text Simplification Tool for Health Content. *AMIA Annual Symposium Proceedings*, *2010*, 366–370.

Keselman, A., Tse, T., Crowell, J., Browne, A., Ngo, L., & Zeng, Q. (2007). Assessing consumer health vocabulary familiarity: an exploratory study. *Journal of Medical Internet Research*, *9*(1), e5.

Landauer, T. K., & Dutnais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *PSYCHOLOGICAL REVIEW*, *104*(2), 211–240.

Laurent, M. R., & Vickers, T. J. (2009). Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association*, *16*(4), 471–479.

Lebret, R., & Collobert, R. (2014). Word Embeddings through Hellinger PCA. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 482–490.

Lerner, E. B., Jehle, D. V., Janicke, D. M., & Moscati, R. M. (2000). Medical communication: do our patients understand? *The American Journal of Emergency Medicine*, *18*(7), 764–766.

Liu, F., Moosavinasab, S., Agarwal, S., Bennett, A. S., & Yu, H. (2013). Automatically identifying health- and clinical-related content in wikipedia. *Studies in Health Technology and Informatics*, *192*, 637–641.

McCrae, J., & Collier, N. (2008). Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC Bioinformatics*, *9*(1), 159.

McInnes, B. T., Pedersen, T., & Carlis, J. (2007). Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain. *AMIA Annual Symposium Proceedings*, *2007*, 533–537.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Mnih, A., & Hinton, G. E. (2009). A Scalable Hierarchical Distributed Language Model. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 1081–1088). Curran Associates, Inc.

Moen, S. P. F. G. H., & Ananiadou, T. S. S. (n.d.). Distributional Semantics Resources for Biomedical Text Processing.

Nazi, K. M., Hogan, T. P., McInnes, D. K., Woods, S. S., & Graham, G. (2013). Evaluating patient access to Electronic Health Records: results from a survey of veterans. *Medical Care*, *51*(3 Suppl 1), S52–56.

Neelakantan, A., & Collins, M. (2015). Learning Dictionaries for Named Entity Recognition using Minimal Supervision. *arXiv:1504.06650 [cs, Stat]*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, *12*.

Plovnick, R. M., & Zeng, Q. T. (2004). Reformulation of consumer health queries with professional terminology: a pilot study. *Journal of Medical Internet Research*, *6*(3), e27.

Polepalli Ramesh, B., Houston, T. K., Brandt, C., Fang, H., & Yu, H. (2013). Improving Patients' Electronic Health Record Comprehension with NoteAid. In *MedInfo* (Vol. 192, pp. 714–718). IOS Press.

Pyper, C., Amery, J., Watson, M., & Crook, C. (2004). Patients' experiences when accessing their on-line electronic patient records in primary care. *The British Journal of General Practice*, *54*(498), 38–43.

Rajagopalan, M. S., Khanna, V. K., Leiter, Y., Stott, M., Showalter, T. N., Dicker, A. P., & Lawrence, Y. R. (2011). Patient-Oriented Cancer Information on the Internet: A Comparison of Wikipedia and a Professionally Maintained Database. *Journal of Oncology Practice*, *7*(5), 319–323.

Reavley, N. J., Mackinnon, A. J., Morgan, A. J., Alvarez-Jimenez, M., Hetrick, S. E., Killackey, E., … Jorm, A. F. (2012). Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources. *Psychological Medicine*, *42*(08), 1753–1762.

Rocchio, J. (1971). Relevance feedback in information retrieval. In *The Smart Retrieval System: Experiments in Automatic Document Processing.* (pp. 313–323).

Socher, R., Huang, E. H., Pennin, J., Manning, C. D., & Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems* (pp. 801–809).

Turney, P. D. (2001). *Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL*.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, *37*(5), 360–363.

Wang, C., Cao, L., & Zhou, B. (2015). Medical Synonym Extraction with Concept Space Models. *arXiv:1506.00528 [cs]*.

Yu, H., & Agichtein, E. (2003). Extracting synonymous gene and protein terms from biological literature. *Bioinformatics (Oxford, England)*, *19 Suppl 1*, i340–349.

Yu, H., Hripcsak, G., & Friedman, C. (2002). Mapping Abbreviations to Full Forms in Biomedical Articles. *Journal of the American Medical Informatics Association*, *9*(3), 262–272.

Zweig, G., & Burges, C. J. C. (2011). *The Microsoft Research Sentence Completion Challenge* (No. MSR-TR-2011-129).