CS691CL - Computational Linguistics: Syntax and Semantics Project Proposal

Nicholas Monath, Niklas Shulze, Klim Zaporojets

March 26, 2014

1 Introduction

There is no shortage of online databases of documents containing valuable information, but there is a need for more tools to organize this data to provide users a more accessible interface than a standard keyword search. Providing solutions to this problem of information overload has been the focus of years of research in information retrieval, natural language processing and machine learning in general. In particular, much work has been done on defining document similarity measures, which determine the relatedness of two documents based on their text content. Typical approaches to this problem are based in word usage statistics and are not able to capture the essence of a text, as the syntactic and semantic relationships of the words are disregarded. In this project, we will make use of dependency parsing and automatic semantic role labeling to extract syntactically and semantically related phrases from text. Using these phrases, we hope to create a more robust document representation and similarity measure than the traditional approaches.

We would like to apply our proposed document similarity measure to the problem of identifying relatedness between scientific research papers. The similarity of scientific research papers is a less studied problem than document similarity in general and a robust measure is in high demand [14]. Such a measure would be useful in both research paper recommender systems and search engines.

The rest of this document is organized as follows: a review of related work done in this field and explanation of how our work differs from this previous research; our proposed approach to defining a document similarity measure; and an explanation of our proposed experiments.

2 Related Work

Document similarity is a much studied problem in the field of information retrieval with a wide range of applications, such as document classification and clustering, searching in large unorganized datasets, and recommendation systems. The problem of document similarity extends beyond the measurement of relatedness of *unstructured* documents containing only text to *structured* documents, which contain hyperlinks and other annotations [19].

The most common approach to unstructured text document similarity is a vector space, statistics based method known as the *bag-of-words* approach. The details of this approach are represented as a foundation for our model in Section 3.1.1. Work in [12] shows the effectiveness of this simple representation on a number of datasets in the problem of document clustering. Another common technique is the representation of documents with a *language model*. First presented in [27], language models make the assumption that the collection of individual terms in a document is a sample from a probability distribution. This approach is particularly useful in unsupervised document clustering. Language models are often made more robust by adding additional information such as underlying topic labels obtained algorithms such as *Latent Dirichlet Allocation* (LDA) or *probabilistic latent semantic analysis*. These

models have been shown to be effective in document retrieval and categorization by [11]. These approaches are also used in our approach and presented in more detail in Section 3.1.1.

In these traditional approaches, the syntactic and semantic relationships of words and phrases in the text are ignored. The models are based on statistical information on the frequency of the occurrence of word sequences or n-grams. Often unigrams (i.e. $n = 1$) are used and so the measure is the frequency of the occurrence of single words. Bigrams (two-word sequences) and trigrams (three-word sequences) are also commonly used. Intuitively, the substitution of *n-grams* with syntactically related groups of words is a logically sound choice. Often, *n-grams* are disparate sequences of terms, while syntactically related groups of words can provide a more robust feature that preserves the semantic meaning of phrases. Initial work was done on this in the late 1990s, such as [9], [6]. The recent works of [24] and [16], show the method can have significant benefits over the traditional *n-gram* approach. Nastase et al in [24] use syntactically related pairs of words obtained using a dependency parser as the base elements in a vector-based bag-of-words approach to perform the task of supervised text classification on the Reuters-21578 dataset. They also experiment with a combination of using the syntactically related pairs and unigrams both with and without syntactic labels. Koster et al in [16] use a similar method of combining syntactically related triples obtained with a dependency parser with unigrams in a vector based model to perform the task of patent document classification.

The most recent work related to document similarity of scientific papers can be found in [13] and [14], it uses a unigram based language model to represent the documents. The model treats the text as *semi-structured*, taking advantage of the additional features of the keywords of the paper, authors' names, and the name of the journal in which the paper appears.

Our work separates itself from these previous studies in that we propose to use syntactically and semantically related sequences of words extracted from the text, using not only dependency parsing, but also automatic predicate argument structure detection, both in a language model and a vector based representation of documents. In creating a language model, we hope to provide a more robust representation of documents using LDA and other techniques. We will also apply the approach to unsupervised document clustering and pairwise document similarity for applications such as recommender systems.

3 Proposed Approach

There are two components of a **document similarity measure**: first, the definition of a *document representation*, which specifies the data structure used to represent the document in a way which is meaningful to a machine; and second, a *distance function*, which assigns to a given pair of documents a real number representing how different the two documents are. We propose several new document similarity measures which incorporate the output of linguistic tools such as dependency parsing and automatic semantic role labelers in traditional statistics-based approaches. To determine the effectiveness a document similarity measure, the measure must be applied to an information retrieval task. We will evaluate our proposed similarity measures with the tasks of **document clustering** and **document classification**.

3.1 Proposed Similarity Measures

3.1.1 Traditional Approaches & Baselines

The following three methods are used as baselines in our experiments. We use a unigram, bigram and trigram version of each approach. These approaches are also used as the foundation for each of our proposed similarity measures presented in Sections 3.1.2, 3.1.3, and 3.1.4.

Bag of Words

In a vector space n-gram bag of words approach, the *document representation* is a feature vector of length M, where M is the number of unique n-grams in the entire collection of documents. Each element of the feature corresponds to a word and value of the feature for each element is typically either binary or the term's tf-idf (term-frequence-inverse-document-frequency) value. Three *distance functions* can be used, the cosine distance, Jaccard Coefficient, and Pearson Correlation Coefficient as presented in [12].

Language Model

Another canonical approach to document similarity is to use a n-gram language model. The *document representation* is the language model, which, for document d_i is defined as:

$$\mathcal{D}_i(w) = \lambda P(w|d_i) + (1-\lambda)P(w|C) \tag{1}$$

where w is an n-gram and C is the entire collection of documents. The *distance function* is the Kullback-Leiber divergence:

$$\mathrm{KL}(\mathcal{D}_i \| \mathcal{D}_j) = \sum_i \ln\left(\frac{\mathcal{D}_i(w)}{\mathcal{D}_j(w)}\right)\mathcal{D}_i(w) \tag{2}$$

Language Model with LDA

In this approach, the n-gram language model is made more robust with the use of Latent Dirichlet Allocation (LDA) [1]. Theoretically, LDA will allow us to discover latent topics in the documents, and further improve the language model of an article based on the revealed topical information. The *document representation* is now defined as:

$$\mathcal{D}_i(w) = \lambda_1 P(w|d_i) + \lambda_2 P(w|C) + \lambda_3 P(w|T) \tag{3}$$

where T is the set of topics underlying document d_i as defined by LDA. The *distance function* is still the Kullback-Leiber divergence.

3.1.2 Making use of Dependency Relations

A dependency parser such as the *Stanford Parser* [35] can be used to extract the inter word dependencies in a sentence. For example, given the sentence:

The quick brown fox jumped over the lazy dog

The outputted dependencies are:

```
det(fox-4, The-1)          amod(fox-4, quick-2)
amod(fox-4, brown-3)       nsubj(jumped-5, fox-4)
root(ROOT-0, jumped-5)     det(dog-9, the-7)
amod(dog-9, lazy-8)        prep_over(jumped-5, dog-9)
```

The word pairs of each dependency relation are used in place of n-grams in the *document representations* of the three document similarity measures described in Section 3.1.1. We will experiment with including the type of relationship along with the word pairs. Also, we will experiment with using the dependency pairs along with the unigrams in the three approaches.

3.1.3 Making use of Predicate-Argument Structure

A automatic semantic role labeler such as the *Illinois Semantic Role Labeler (SRL)* [29] can be used to determine the predicate argument structure of English sentences. For example, given the sentence:

`After eating dinner, the quick brown fox saw the lazy dog, who was still sleeping.`

The following predicate argument relationships are given by the Illinois Semantic Role Labeler:

Relation #1:	eat.01	meal [A1]	consumer/eater [A0]	
	`eating`	`dinner`	`the quick brown fox`	
Relation #2:	see.01	viewer [A0]	thing viewed [A1]	temporal [AM-TMP]
	`saw`	`the quick brown fox`	`the lazy dog`	`after eating dinner`
Relation #3:	sleep.01	sleeper [A0]	sleeper [R-A0]	temporal [AM-TMP]
	`sleeping`	`the lady dog`	`who`	`still`

The groups of terms making up the relations and arguments are used in place of the n-grams in the *document representations* of the three document similarity measures described in Section 3.1.1. We will experiment with including the role labels along with each group of terms. Also, we will try using unigrams in addition to these groups of terms in the three approaches.

3.1.4 Making use of Word2Vec

Vector space word models are useful in determining the syntactic and semantic relatedness between words. In these models, each word in a corpus is assigned a position in a high-dimensional space $\mathbb{R}^N$ through a training process. The hope is that words that are related to one another are placed at a near by in $\mathbb{R}^N$. The relatedness (or rather *difference*) between two words is typically measured by the cosine distance between their two vectors. Word2Vec is a state of the art vector space word model and training process [21]. We propose two modifications to the previously described models that make use of a vector space word model.

The first method is to perform clustering of the words and to use the centroid of each cluster in place of n-grams in the previously described approaches. Specifically, we let $\mathbf{W}$ be the set of words that appear in one or more of the collection of documents $d_1, d_2, \ldots, d_n$ and we let $\mathbf{V}$ be the set of vectors corresponding to the entries in $\mathbf{W}$. We cluster $\mathbf{V}$ using an algorithm such as k-means or DBSCAN and so each entry in $\mathbf{V}$ is associated with a specific cluster. We then calculate the center point or centroid of that cluster. Then each word in the corpus is replaced with the centroid of its associated cluster and the three approaches described in Section 3.1.1 are used.

A second method, but related method, is a less strong form of clustering. Clustering is done in just two passes through the documents. For each word, we find the closest k words in the vector space defined by Word2Vec. If the word is not already associated with a cluster and if any of these k words appear in the collection of documents, a cluster is defined for the related words. If any of the k words appear in other clusters, the clusters are merged together. We then calculate the center point or centroid of that cluster. Then each word in the corpus is replaced with the centroid of its associated cluster and the three approaches described in Section 3.1.1 are used.

According to [21], we can define multiple word phrases in vector space as the sum of the vectors of the individual words. In so doing we can apply these two methods to the multi-word phrases extracted from the dependency parsing and semantic role labeling.

3.1.5 Text Processing Tools

Many of the current papers convert words to their lemmatized form that is removing inflections so that each word is in its base form. This helps to reduce the feature space and make the feature vectors (or

probability distributions) more descriptive. This process will take place after the dependency parsing and semantic role labeling. We can experiment with whether or not it will be used in the Word2Vec model.

3.2 Experiments

One of the main goals of this project is to provide a tool for the organization and search of collections of scientific papers. Unfortunately, there is a limited number of existing datasets of research papers that can be used to evaluate the quality of a document similarity measure on such a domain. We are in the process of obtaining a data set of research papers tagged with meta data for evaluation from the authors of [14]. Additionally, we will perform experiments with our document similarity measures on other datasets such as the Reuters-21578 corpus. These experiments will be in both document clustering and document classification.

3.2.1 Document Clustering

The problem of **document clustering** is defined as the automatic organization of documents into logical groups based on latent features. For example, given a corpus of computational linguistics research papers, we might hope that a clustering of the papers would place all the work on grammars and language representations in one cluster, the work on semantics in another, the work in morphology in a third, etc.

We will reproduce the experiment in [14], which uses a corpus of 209 artificial intelligence research papers manually tagged by experts. For each paper, the 30 papers considered most similar using the bag-of-words approach [1] with the cosine distance function were labeled by an expert as similar or dissimilar. A document similarity measure is evaluated by selecting, for each paper, the 30 closest papers and measure the Mean Average Precision and Mean Reciprocal Rank. We will the results of our approach to that of the authors.

We will also evaluate our document similarity measures on the Reuters-21578 corpus. Each news article in the Reuters corpus is tagged with a class label. We will perform an experiment in which we cluster the articles in the corpus and measure how well each cluster encompasses a class of article and how well each class is represented by a cluster, both in terms of precision and recall. A comparable experiment is done in [11].

Finally, if time allows, we will perform an additional experiment on the clustering of research papers, in which we collect a corpus and for each paper return the top k most similar papers. The list of similar papers will then be evaluated by a human subject and the rate of dissimilar papers will be evaluated.

3.2.2 Document Classification

The problem of **document classification** is formally defined just like the general problem of classification in machine learning: Given input pairs of documents and class labels, $(d_1, c_1), (d_2, c_2), \ldots, (d_N, c_N)$, with $d_1, \ldots d_n \in \mathcal{D}$ and $c_1, \ldots, c_N \in \mathcal{C}$, where $\mathcal{D}$ is the collection of training documents and $\mathcal{C}$ is the set of class labels, assume there is a function f which maps any document d the correct class label c. The goal of learning is to approximate the function f with $\hat{f}$ a function created with the observed training data.

The authors of [24] present results from a document classification experiment on the Reuters-21578 dataset using a vector space model which makes use of dependency parsing. We will repeat their experiment to see how our model performs in comparison to theirs. The classification experiment is done using the 10 most frequent classes in the Reuters-21578 document collection with the ModApte split designating the specific documents for training and testing. The precision and recall scores are used to judge the performance of a document similarity measure.

[1] With tf-idf as the value of the vector for each unigram

3.2.3 Clustering and Classification Methods

We will use a standard set of machine learning tools to implement the clustering and classification procedures. For clustering, we will perform experiments using *k-means*, *DBSCAN*, and the *EM Algorithm*. For classification, we will perform experiments using *Support Vector Machines*, *Naive Bayes*, and *k-Nearest Neighbor* approaches. We will be programming most of our project in `Python` and will be using the machine learning tools available in `NumPy` and `SciPy`.

Note the below references includes papers that we read, but did not cite in this document. Those papers are listed for our own reference.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] Alexander Budanitsky and Graeme Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2, 2001.

[3] Poonam Chahal, Manjeet Singh, and Suresh Kumar. Ranking of web documents using semantic similarity. In *Information Systems and Computer Networks (ISCON), 2013 International Conference on*, pages 145–150. IEEE, 2013.

[4] Hung Chim and Xiaotie Deng. A new suffix tree similarity measure for document clustering. In *Proceedings of the 16th international conference on World Wide Web*, pages 121–130. ACM, 2007.

[5] Jon D. McAuliffe David M. Blei. Supervised topic models. *The Scientific World Journal*.

[6] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155. ACM, 1998.

[7] Tamer Elsayed, Jimmy Lin, and Douglas W Oard. Pairwise document similarity in large collections with mapreduce. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 265–268. Association for Computational Linguistics, 2008.

[8] Samuel Fernando, Mark Hall, Eneko Agirre, Aitor Soroa, Paul Clough, and Mark Stevenson. Comparing taxonomies for organising collections of documents. In *COLING*, pages 879–894, 2012.

[9] Johannes Furnkranz, Tom Mitchell, and Ellen Riloff. A case study in using linguistic phrases for text categorization on the www. In *Proceedings from the AAAI/ICML Workshop on Learning for Text Categorization*, pages 5–12, 1998.

[10] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. Semantic measures for the comparison of units of language, concepts or entities from text and knowledge base analysis. *arXiv preprint arXiv:1310.1285*, 2013.

[11] Thomas Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. 2000.

[12] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.

[13] German Hurtado Martin, Steven Schockaert, Chris Cornelis, and Helga Naessens. Finding similar research papers using language models. In *2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation (SPIM-2011)*, pages 106–113. University College Ghent, 2011.

[14] GermáN Hurtado MartíN, Steven Schockaert, Chris Cornelis, and Helga Naessens. Using semi-structured data for assessing research paper similarity. *Information Sciences*, 221:245–261, 2013.

[15] Elias Iosif and Alexandros Potamianos. Unsupervised semantic similarity computation between terms using web documents. *Knowledge and Data Engineering, IEEE Transactions on*, 22(11):1637–1647, 2010.

[16] Cornelis HA Koster and Jean G Beney. Phrase-based document categorization revisited. In *Proceedings of the 2nd international workshop on Patent information retrieval*, pages 49–56. ACM, 2009.

[17] Ainura Madylova and SG Oguducu. A taxonomy based semantic similarity of documents using the cosine measure. In *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on*, pages 129–134. IEEE, 2009.

[18] Ana G Maguitman, Filippo Menczer, Heather Roinestad, and Alessandro Vespignani. Algorithmic detection of semantic similarity. In *Proceedings of the 14th international conference on World Wide Web*, pages 107–116. ACM, 2005.

[19] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[20] Olena Medelyan, Steve Manion, Jeen Broekstra, Anna Divoli, Anna-Lan Huang, and Ian H Witten. Constructing a focused taxonomy from a document collection. In *The Semantic Web: Semantics and Big Data*, pages 367–381. Springer, 2013.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[23] Mandar Mitra, Chris Buckley, Amit Singhal, Claire Cardie, et al. An analysis of statistical and syntactic phrases. In *RIAO*, volume 97, pages 200–214, 1997.

[24] Vivi Nastase, Jelber Sayyad, and Maria Fernanda Caropreso. Using dependency relations for text classification. *University of Ottawa SITE Technical Report TR-2007-12*, 13, 2007.

[25] Alexandre Passos and Jacques Wainer. Wordnet-based metrics do not seem to help document clustering. In *Proc. of the of the II Workshop on Web and Text Intelligente, São Carlos, Brazil*, 2009.

[26] Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299, 2007.

[27] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[28] Andrei Popescu-Belis and Majid Yazdani. Using a wikipedia-based semantic relatedness measure for document clustering. In *Graph-based Methods for Natural Language Processing*, number EPFL-CONF-167425, 2011.

[29] Vasin Punyakanok, Dan Roth, and Wen-tau Yih. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287, 2008.

[30] Muhammad Rafi and Mohammad Shahid Shaikh. An improved semantic similarity measure for document clustering based on topic maps. *arXiv preprint arXiv:1303.4087*, 2013.

[31] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[32] Jaz Kandola John Shawe-Taylor and Nello Cristianini. Learning semantic similarity. 2002.

[33] Sheetal A Takale and Sushma S Nandgaonkar. Measuring semantic similarity between words using web documents. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 1(4), 2010.

[34] Nguyen Chi Thanh and Koichi Yamada. Document representation and clustering with wordnet based similarity rough set model. *International Journal of Computer Science Issues (IJCSI)*, 8(5), 2011.

[35] Stanford CoreNLP Tools. The stanford natural language processing group.

[36] Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16. ACM, 2005.

[37] Andrea Varga, Amparo Elizabeth Cano, and Fabio Ciravegna. Exploring the similarity between social knowledge sources and twitter for cross-domain topic classification. In *Proceedings of the Knowledge Extraction and Consolidation from Social Media, 11th International Semantic Web Conference (ISWC2012)*, 2012.

[38] John S Whissell and Charles LA Clarke. Effective measures for inter-document similarity. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1361–1370. ACM, 2013.

[39] Seok-Ho Yoon, Ji-Soo Kim, Jiwoon Ha, Sang-Wook Kim, Minsoo Ryu, and Ho-Jin Choi. Link-based similarity measures using reachability vectors. *The Scientific World Journal*, 2014, 2014.

[40] Ming Zhang, Weichun Wang, and Xiaoming Li. A paper recommender for scientific literatures based on semantic concept similarity. In *Digital Libraries: Universal and Ubiquitous Access to Information*, pages 359–362. Springer, 2008.

[41] Zhicheng Zheng, Xiance Si, Fangtao Li, Edward Y Chang, and Xiaoyan Zhu. Entity disambiguation with freebase. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 82–89. IEEE Computer Society, 2012.

[42] Jianhan Zhu, Xiangji Huang, Dawei Song, and Stefan Rüger. Integrating multiple document features in language models for expert finding. *Knowledge and Information Systems*, 23(1):29–54, 2010.