Related Work Search using Frame-Semantic Parsing

Abstract

Searching scientific literature is typically a matter of querying search engines with scientific keywords that come to mind as reflecting one's interests. However, often a user's point of departure is another scientific paper or article, e.g., a paper that the user authored or is writing an essay about. Consequently, the literature search is for related work that may support or contradict the hypotheses or the results in the first paper. This study presents an approach to automating the search for related literature based on frame-semantic parsing for extracting hypotheses and results. We test our models on 40 random articles and show that our frame-semantic model retrieves related documents with more than 75% precision.

1 Introduction

Students and researchers spend a lot of their time looking for related work. The body of research literature grows faster and faster, and it is hard to keep up with recent developments. When reading or writing a paper, students and researchers have to identify the main assertions of that paper and find all or most related research that support, argue against or contradict these assertions. This is a core part of academic scholarship, pushing the horizons of science, standing on the shoulders of giants; not reinventing the wheel.

The task of finding related work, however, is very time-consuming, and it becomes increasingly hard to exhaust the search space of possible references. It is therefore necessary to, at least partially, automate the task.

So far, the task has been relieved a bit in some scientific domains by information retrieval (IR) tools, such as The ACL Anthology Searchbench (Schäfer et al. 2011), often using NLP techniques to make the search more targeted than standard web search. However, these tools still take keywords as input and use search filters as guidance. We want to go considerably beyond this.

Our aim, in contrast, is to automate the search for related literature with frame-semantic parsing and use this to deliver structured search results – such as in Figure 1 – that also inform end users *why* (or in what way) related work is related; initially, with an application to the ACL Anthology Reference Corpus (Bird et al. 2008). Frame semantic parsing has previously been used for knowledge extraction (Søgaard, Plank, and Alonso 2015). Our approach can be divided into three steps: Given a reference paper, the first step involves the use of frame-semantic parsing to limit the search space from nearly 16,000 candidates to the top 2,000 most closely related candidates. In the second step, these candidates are ranked according to the number of bigrams and trigrams they share with the reference paper. Finally, the third step re-ranks the top 50 candidates, using frame-semantics to compare the reference paper and the candidate related work, and figure out *how*, if at all, they are related.

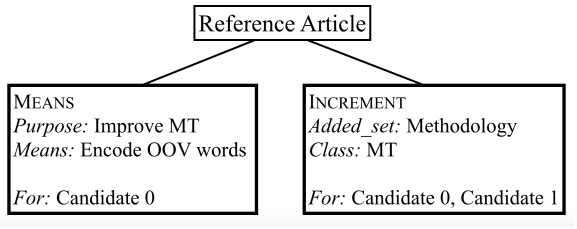

Figure 1: Structured related work search results. Candidates 0 and 1 are scientific papers found to be related to the reference article.

Figure 2 provides excerpts from the abstracts of two research papers that were identified to be related by the system presented below, because they both work on how to handle rare or previously unseen words in the context of neural machine translation. The reference is an ACL 2015 paper; however, candidate 0 is an EMNLP 2015 paper, published only two months apart. The candidate should have cited the reference in this specific case since their model is much more generic. However, we understand that it is even harder to find related work when the two papers have been published in such a short time apart from each other. This specific point makes the existence of a system like ours even more important, since our model is able to find related work from a large repository in a matter of seconds.

Contributions We are, to the best of our knowledge, the first to introduce the task of structured retrieval of related scientific papers or articles, informing end users why suggested papers were found to be relevant. We show that

Reference

Variable-Length Word Encodings for Neural Translation Models

Recent work in neural machine translation has shown promising performance, but the most effective architectures do not scale naturally to large vocabulary sizes. **We propose and compare three variable-length encoding schemes that represent a large vocabulary corpus using a much smaller vocabulary with no loss in information.** Common words are unaffected by our encoding, **but rare words are encoded using a sequence of two pseudo-words.** Our method is simple and effective: it requires no complete dictionaries, learning procedures, increased training time, changes to the model, or new parameters. **Compared to a baseline that replaces all rare words with an unknown word symbol**, our best variable-length encoding strategy improves **WMT English-French** translation performance by up to 1.7 BLEU.

Candidate 0

Addressing the Rare Word Problem in Neural Machine Translation

Neural Machine Translation (NMT) is a new approach to machine translation that has shown promising results that are comparable to traditional approaches. A significant weakness in conventional NMT systems is their inability to correctly trans-late very rare words: **end-to-end NMTs tend to have relatively small vocabularies with a single unk symbol that represents every possible out-of-vocabulary** (OOV)word. In this paper, **we propose and implement an effective technique to address this problem.** We train an NMT system on data that is augmented by the output of a word alignment algorithm, **allowing the NMT system to emit, for each OOV word in the target sentence, the position of its corresponding word in the source sentence. This information is later utilized in a post-processing step that translates every OOV word using a dictionary**. Our experiments on the **WMT'14 English to French** translation task show that this method pro-vides a substantial improvement of up to2.8 BLEU points over an equivalent NMT system that does not use this technique. With 37.5 BLEU points, our NMT system is the first to surpass the best result achieved on a WMT'14 contest task.

Figure 2: An example of a candidate article which has been identified by our approach to be relevant to the reference article. Both scientific papers tackle the problem of rare/unknown words in Neural Machine Translation and are highly relevant to each other.

ID	Causation	Increment	Means	Aggregate
P15-2068	18	18	15	12
P86-1035	15	26	11	15
A00-1012	17	12	12	4
P93-1256	11	22	3	10
D56-1984	9	19	13	4

Table 1: We represent scientific papers by frame distributions; see five example papers above. P15-2068 and P86-1035 are similar in terms of the distribution of the most common frames in each article.

frame-semantic parsing can be used to provide this level of structure in the output, but also to improve our related work search considerably.

2 Model description

Our approach to related work search is a three-step approach. Our first step is a crude, coarse-grained preselection. For that, as well as the final step of our approach, we use frame-semantic parsing. The overall algorithm for related work search, that we propose, is presented in Algorithm 1. See the related work section for background on frame-semantic parsing.

2.1 Frame semantics based selection

Our version of the ACL Anthology Reference Corpus contains 15,819 articles. In Step 1, we aim to reduce this search space to a smaller set of potential candidates for retrieval. We want to filter out every paper that is not in the semantic space of the reference paper. In order to assess if two papers are in the same semantic space, we look at their distribution over semantic frames, by doing frame-semantic parsing of the contents of the papers, and compute their similarity (distance) in a semantic space of conceptual frames.

This method is relatively naïve, but fast and efficient. We start by counting the number of unique frames in every paper and store the result in a vector. Table 1 shows a subset of the most recurrent frames that we have extracted during this step.

Semantic parsing is of course relatively costly, but this can be done off-line. Additionally, the indexation of papers by semantic frames is more compact than keyword-based indexation. Once the data-frame is built, the similarity in terms of *frame overlaps* between two papers is given by the dot product of the two corresponding lines in the data-frame seen as vectors; any other distance metric in the frame vector space can be easily computed. Then, given a reference article, dot products (or any other distance metric) provide a ranking of all the candidate papers. In our naïve implementation, computing this global ranking takes 3–5 seconds.

Algorithm 1 Related work search algorithm

```
// hash table with frames as keys
//  and important sentences as
//  values for the reference :
reference_paper = hash_table(reference_paper)
// List of the most significant frames :
frames ← [Causation, Increment, Mean,...]
for each article do
    SEMAFOR(article)
    candidate_papers ← similar papers with semantic frames counts and n-grams
    for each candidate_paper do
        score ← 0
        // hash table with frames as keys
        //  and important sentences as
        //  values for each candidate :
        candidate_paper = hash_table(candidate_paper)
        for each frame ∈ frames do
            for each ref_sentence ∈ reference_paper[frame] do
                for each cand_sentence ∈ candidate_paper[frame] do
                    common_bigrams ← ref_sentence ∩ cand_sentence
                    for each bigram ∈ common_bigrams do
                        tf_idf_cand ← tf_idf(bigram, cand)
                        tf_idf_ref ← tf_idf(bigram, ref)
                        score += tf_idf_ref * tf_idf_cand
                    end for
                    total_score += score / |common_bigrams|
                end for
            end for
        end for
    end for
    // Final score given to a candidate :
    total_score = total_score / number of comparisons
end for
```

2.2 *n*-gram based selection

The distribution of frames provides information about the way a document is semantically structured, but not about the exact claims of scientific papers. While our frame-semantic parser also provides argument annotations for the frames, which we will use to provide structured search results back to the end user, the argument annotations are error prone, and we instead propose to use n-gram features for more fine-grained retrieval of related work. This part of our related work search engine is a relatively standard IR technique.

Specifically, we compute document similarity using TF-IDF bag-of-words representations with bigrams and trigrams. For the purpose of related work search, we extract the 50 most important bigrams and trigrams for each article. We then rank the 2,000 articles according to how many bigrams and trigrams they share with the reference article. Finally, we keep the 50 highest ranked articles as candidates for the last step. Bigrams and trigrams with high TF-IDF values were found to supplement the frame-level information well, giving us a more fine-grained picture of how papers are related.

We note that word embeddings trained on scientific literature were not useful in this step because they assign similar representations to words with similar distributions, and while the phrase *Perceptron-Based* has a similar distribution as *Naive Bayes-based* or *NN-based* – in a context of finding related scientific papers, these can be exactly the distinctions that we are looking for.

2.3 Ranking

Our next step is to rank the 50 obtained articles according to their similarity with the reference paper. The purpose is to be as accurate as possible; we therefore combine the frame-semantic parser output with the bigrams, trigrams and their respective TF-IDF values. After that, we iterate through all 50 articles and do pairwise comparisons of sentences in the two articles that contain the same semantic frame.

In the following section, we map every reference article with each candidate through a hash table, with the keys being the semantic frame names (*e.g: Causation, Likelihood, Cause_to_make_progress*). Their values consists of the sentences from the article that contain one instance of this frame type and a mean TF-IDF value greater than a threshold (which was set experimentally); this representation is what we use for the ranking step.

While ranking, we successively compare the sentences containing a frame type with each sentence that contains the same frame type in the reference article. A score between 0 and 1 is assigned to all of those comparisons, where the final score given to a candidate is the mean of all the comparisons performed between the sentences of the reference article and the candidate. The candidate with a **high score** is then selected, it represents the article which shared the most "frame-specific sentences" with the reference article.

Finally, we have computed the sentence-to-sentence compatibility; we first turn the two sentences into their respective list of bigrams, and we then extract in a third list called common_bigrams the ones which appear in both. Then, if this list

is not empty, we assign to this specific comparison the value given on line 21 of the algorithm, which combines the TF-IDF values of the common bigrams from both the reference and the candidate paper.

We repeat that atomic comparison step between every pair of sentences containing both the same frame type, one being from the candidate and the other from the reference paper. The global score assigned to a candidate is the mean of all those sentence-to-sentence comparisons.

Finally, we calculate how much two sentences are close to each other by scoring how many sharing some *n*-grams that are important to both texts, the two sentences containing an occurrence of the same frame type. We tried to add the trigrams into that part of the process, but it was so rare to have two sentences containing at the same time the same frame type and some trigrams in common that it was basically just adding some more computation time without giving better results.

We test two different models, which we call FRAME SEMANTICS and FRAME SEMANTICS++. The two models are almost similar except for one minor difference in the last step. Our FRAME SEMANTICS model bases the comparison between articles on solely the comparison of the frame elements in the shared frames between the reference and the candidate. The FRAME SEMANTICS++ model on the other hand, does not only compare the frame elements, but the full frame-semantic parses, comparing all sentences in the reference article and the candidate paper in a pair-wise fashion. Algorithm 1 contains the pseudocode for FRAME SEMANTICS++. Our implementation, as well as the preprocessed data used in our experiments, will be made available upon acceptance.[1]

3 Experiment Setup

3.1 Dataset

To evaluate our approach, we use The ACL Anthology Reference Corpus. The ACL Anthology Corpus is a repository of the papers accepted at ACL conferences and in ACL journals (including papers from Computational Linguistics, Transactions of the ACL, ACL, EACL, NAACL, EMNLP, and CoNLL), until 2015. The corpus is meant to be used in *benchmarking applications for scholarly and bibliometric data processing*. It is publicly available online (`http://acl-arc.comp.nus.edu.sg/`). We retrieved 22,878 articles in XML format before preprocessing.

We split our corpus into a training set and a test set. Our TF-IDF vectorization, for example, is fit on the training set, consisting of the majority of the data (15,819 articles).[2] The test set, composed of 40 articles, is randomly selected from the remaining articles, to assess how good our proposed models generalize to unseen data.

[1] http://github.com/anonymized

[2] We had to throw out about 7,000 articles that either turned out to not be scientific articles, but introductions written by editors or short academic abstracts, or were not in well-formatted XML.

	Inter-annotator agreement
FRAME SEM.	
κ	**59.23**
A_obs	75.00
A_exp	61.5
FRAME SEM.++	
κ	**65.83**
A_obs	83.00
A_exp	52.50

Table 2: Inter-annotator agreement for all candidates for both FRAME SEMANTICS and FRAME SEMANTICS++. We note that there is a high agreement between the annotators since $\kappa > 50\%$ on all cases. A_obs is the observed agreement, whereas A_exp is the expected agreement.

	FRAME SEM.	FRAME SEM.++
Relevant	**78.3%**	69.16%
Highly Relevant	43.33%	**44.16%**

Table 3: Percentage of relevancy per candidate-reference pair. Relevant means that there was at least one annotator which has voted relevant. Highly relevant means that all annotators have voted relevant.

3.2 Evaluation

Human judgments To evaluate our model, we hired three human annotators. Each annotator was given 40 references articles, each along with three candidates for related work, as ranked by the FRAME SEMANTICS and FRAME SEMANTICS++ models, i.e., 40 sets of a reference article and a total of six candidates. For all six pairs of reference-candidates in each set, they were asked to judge whether the candidate was indeed related work or not. All the annotators were students of computer science or related fields.

Inter-annotator agreement Human judgments are relatively uncontroversial for evaluating information retrieval systems, but human judges disagree, and we therefore first measure the agreement between our three annotators. We compute inter-annotator agreement using Fleiss' kappa (Fleiss and others 1971), a modified version of Cohen's kappa (Cohen 1960). Table 2 shows the inter-annotator agreement scores.

4 Results

In order to evaluate FRAME SEMANTICS and FRAME SEMANTICS++, we have divided our evaluation into two metrics: Relevant and Highly Relevant. Each annotator had to vote if each tuple candidate-reference were relevant to each other. Given that we tested on 40 completely different articles and that for each article we had 3 candidates, each annotator had 120 candidate-reference to assess. Relevant candidates represent the cases where at least one annotator voted

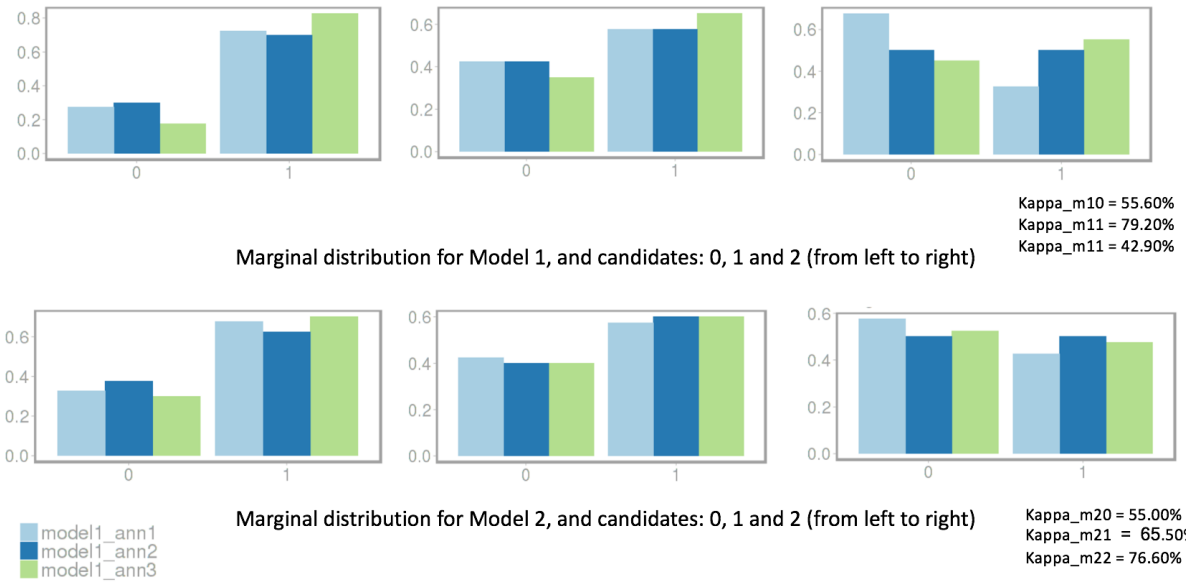

Figure 3: The marginal distribution for the agreement between the three annotators as well as the detailed agreement for each candidate for both models. m_{10} for instance, refers to the first model, and the first candidate.

relevant. However, Highly Relevant refers to cases that were voted relevant by all three annotators.

Our results show that both our methods are relatively good with a low precision of around 70%. Both FRAME SEMANTICS and FRAME SEMANTICS++ achieve a high percentage of 44% on the "Highly Relevant" measure. These results show that, a semantic frame analysis provides a good a way to start document matching in general. Except for running SEMAFOR in the beginning, we have proven that our new approaches are both fast and efficient given the small amount of data we have.

5 Examples and Analysis

To give a more accurate interpretation of our results, we thoroughly analyzed the positive and negative examples. We have noted that the proposed model is relatively good when the input is dealing with recurrent topics in NLP, such as "Word Sense Disambiguation", "Neural Machine Translation", "Context Free Grammars" or "Speech Processing". Almost 100% of the time, the model outputted 3 related candidates, which generally do not only capture the topic, but they also capture other dimensions of the text, such as methods employed or datasets used in the experiments.

We have depicted below some examples extracted from our test set. Examples (1) and (2) are about "Latent Semantics" and "Phrase based Machine Translation", respectively. Example (2) is very good since our model was able to output a candidate which is also related to decoding phrase-based machine translation models. Example (1) is also excellent since all of the candidates are highly correlated and they discuss the exact same topic.

(1) REF Learning the **Latent Semantics** of a Concept from its Definition
OUT A Simple Unsupervised **Latent Semantic**s based Approach for Sentence Similarity
OUT Modeling **Sentences** in the **Latent Space**

(2) REF Entropy-based **Pruning** for **Phrase-based Machine Translation**
OUT Integrating **Phrase-based** Reordering Features into a **Chart-based Decoder for Machine Translation**

However, our approach showed negative results when the input was on a more unusual topic such as "Emails Summarization" or "Sarcasm Detection". In these cases, the expected output would have been articles using at some point the same methods or dealing with the closest available topic in the dataset. Instead, we often obtained a noisy output, *i.e* some articles dealing with totally different subjects, or even some articles that were noisy themselves.

Indeed, even after filtering out noisy articles, the set of preprocessed articles still contained a few papers that either were not scientific papers or which had not been properly handled by the preprocessing. Another special case was to retrieve different or rewritten versions of some articles we used as input. In fact, our algorithm was always capable of retrieving such pairs of almost "twin" articles, probably thanks to the use of bigrams and trigrams.

Examples below are an instance of the few negative examples that our system outputted. Example (3) is confusing for the system because there are not that many articles that tackle "Realization Ranking". Therefore, our model output was very close to the subject since it uses "Minimum Dependency Length", the same method the reference article uses

but with "Grammars Optimization". In example (4), however, the output article was completely different from the reference. This could be explained by the fact that the reference article is about "Summarizing Neonatal Time Series Data", which is an uncommon topic in NLP in general and in the ACL Anthology corpus in particular.

(3) REF Minimal Dependency Length in Realization Ranking
OUT Optimizing Grammars for Minimum Dependency Length
OUT Minimal-length linearizations for mildly context-sensitive dependency trees

(4) REF Summarizing Neonatal Time Series Data
OUT Minimally-Supervised Extraction of Entities from Text Advertisements

(5) REF **Lexicalized Context-Free Grammars**
OUT Capturing **CFLs with Tree Adjoining Grammars**

Example 5 shows a very good case where the candidate article is a follow-up work of the reference article. Both articles are about Lexicalized Context Free Grammars where the reference paper is the original paper defining Lexicalized CFGs. However, the candidate proposed by our algorithm is a follow-up work which defines an algorithm that parses LCFG in a cubical time by allowing less restrictions than the reference paper.

6 Related Work

Frame semantics, introduced by Charles Fillmore (Fillmore 1982), is a formalism for representing our knowledge about conceptual frames, such as CAUSATION, and how these frames are expressed in natural language (Baker, Fillmore, and Lowe 1998). In the FrameNet project,[3] more than 1,200 frames are listed, accompanied with examples, usual trigger words, and common argument roles, encoding knowledge about likely participants in the conceptual frames. Hence, a frame may be defined as background knowledge connecting groups of words (Jurafsky and Martin. 2018). In our study, the frame-semantic parser SEMAFOR (Das et al. 2010; Das et al. 2014) is used to identify and compare meaningful sentences related to hypotheses and results in the scientific papers. SEMAFOR is trained on the FrameNet lexicon. For each parsed sentence, frames are evoked by specific words, and the frames, in turn, define frame-specific semantic roles referred to as *frame elements*.

For instance, a sentence such as "We improved the model by using a neural network" triggers a frame called **Cause_to_make_progress** with "We" corresponding to the frame element **Agent** and "the model" corresponding to the frame element **Project**. Frame names are unique. Some frame elements or roles are optional. For a quick overview of SEMAFOR, an online demo is available at `http://demo.ark.cs.cmu.edu/parse?`.

[3] https://framenet2.icsi.berkeley.edu/

Recently, there has been some work on relation extraction for scientific literature. SemEval 2017 Task 10,[4] for example, was about extracting keyphrases and relations from scientific publications, for example. Most of the extracted keyphrases and relations in this task were not relevant for related work search, however. This is why we explicitly focus on results and hypothesis in this work. We believe, however, that the best models participating in this shared task could likely be used in an ensemble with semantic parsers to extract even better frames and relations from scientific literature. The best systems in the SemEval 2017 Task 10 relied on recurrent neural networks (Augenstein et al. 2017).

7 Conclusion

We have presented a new semantic frame-based model for finding related works in collections of scientific literature. Our retrieval system takes a scientific reference article as input, rather than a query or a list of keywords, analyzes it to find results and hypotheses, and retrieves other scientific papers that support or contradict these hypotheses or results. Our frame-semantic retrieval system enables us to deliver structured output, informing the end user *why* related work was deemed relevant. Our retrieval system was shown to be very precise in experiments with human judges, where our system was shown to be able to retrieve three papers by reference article of which, on average, 3/4 were deemed relevant by at least one annotator, and about half were deemed relevant by all annotators.

References

[Augenstein et al. 2017] Augenstein, I.; Das, M.; Riedel, S.; Vikraman, L.; and McCallum, A. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *SemEval*.

[Baker, Fillmore, and Lowe 1998] Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, 86–90. Stroudsburg, PA, USA: Association for Computational Linguistics.

[Bird et al. 2008] Bird, S.; Dale, R.; Dorr, B. J.; Gibson, B. R.; Joseph, M. T.; Kan, M.-Y.; Lee, D.; Powley, B.; Radev, D. R.; and Tan, Y. F. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*.

[Cohen 1960] Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1):37.

[Das et al. 2010] Das, D.; Schneider, N.; Chen, D.; and Smith, N. A. 2010. Semafor 1 . 0 : A probabilistic frame-semantic parser.

[Das et al. 2014] Das, D.; Chen, D.; Martins, A. F. T.; Schneider, N.; and Smith, N. A. 2014. Frame-semantic parsing. *Computational Linguistics* 40(1):9–56.

[4] https://scienceie.github.io/

[Fillmore 1982] Fillmore, C. J. 1982. Frame semantics. In *Linguistics in the Morning Calm*.

[Fleiss and others 1971] Fleiss, J., et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

[Jurafsky and Martin. 2018] Jurafsky, D., and Martin., J. H. 2018. *Semantic Role Labelling*. Draft of august 15, 3rd edition. chapter 18.

[Schäfer et al. 2011] Schäfer, U.; Kiefer, B.; Spurk, C.; Steffen, J.; and Wang, R. 2011. The acl anthology searchbench. In *ACL*.

[Søgaard, Plank, and Alonso 2015] Søgaard, A.; Plank, B.; and Alonso, H. M. 2015. Using frame semantics for knowledge extraction from twitter. In *AAAI*.