A Discriminative Approach to Evidence Linking and Discourse Prediction

Chris Tanner
Brown University
115 Waterman Street
Providence, RI 02912, USA
christanner@cs.brown.edu

Byron C Wallace
University of Texas at Austin
1616 Guadalupe
Austin, TX 78701, USA
byron.wallace@utexas.edu

Stephen H Chen
University of Texas at Austin
2110 Speedway
Austin, TX 78705, USA
stephen.chen@utexas.edu

Eugene Charniak
Brown University
115 Waterman Street
Providence, RI 02912, USA
ec@cs.brown.edu

Abstract

Text documents of varying nature (e.g., summary documents written by analysts or published, scientific articles) often cite others as a means of providing evidence to support a claim, attributing credit, or referring the reader to related work. We address two new tasks: (1) evidence linking — given a sentence which cites a source paper, predict the most relevant pieces of text within the said source (i.e., the text which warrants the citation); and (2) discourse prediction — predict the *type* of content that is implicitly cited (e.g., hypothesis, method, results, etc). We detail our initial discriminative-based experiments and results, lessons learned, and ideas for future work.

1 Introduction

The amount of online documents continues to grow at an astonishing rate, partly due to the ease to which users can generate content (e.g., social media, Web 2.0, Wiki* articles). Further, many different types of documents link or cite other documents (e.g., websites, research articles, analyst summary reports), and for various reasons: to attribute credit, provide evidence, refer the reader to related work, etc. Due to there being an abundance of documents which cite one another, it could be highly useful to have a system which, given a sentence which cites a document, could automatically return the most relevant sentence(s) from the *cited* document. We refer to this task as *evidence linking*. The sentence within the *citing* document which has the citation is referred to as the *citance.*

In addition, we also aim to predict the *type* of evidence each citation represents. The five pre-defined options follow:

- Hypothesis Citation
- Method Citation
- Results Citation
- Implication Citation
- Discussion Citation

These two tasks, *evidence linking* and *discourse prediction*, were constructed by NIST and comprised their TAC 2014 Biomedical Summarization Workshop.

We describe related work in Section 2, along with our methods in Section 3. Our experiments and results are detailed in Section 4. We quickly discuss our ideas in Section 5, and we conclude our work in Section 6.

2 Related Work

At a more broad scope, many have researched the task of *citation prediction* – that is, given only a document's text (we will refer to this document as the *report*), try to predict which documents it cites as sources (selecting from a large corpus of candidate sources). The seminal work in this task started with Hofman and Cohn's PHITS system [4] which, based on probabilistic latent semantic analysis [5], predicted citations based on documents' (i.e., candidate report-source pairs) similarity in topical content. Extending this work, Erosheva et. al. [2] replaced PLSA with LDA [1] as the fundamental generative process. Later, others researched alternative LDA-based, generative model [10] [7] [8] – all of which, at testing time, determines if a document (report) cites another particular document (source) by sampling from a random variable (often a Bernoulli distribution), which is parameterized by the topic distributions. These systems are agnostic to the actual *citances* – the sentences which contain the citation – and do not assume knowledge of which sentences within the document/report will cite other documents. Assuming such knowledge was considered in the research by Huang et. al. [6], Kataria, et. al. [7], and He et. al. [3]. Again, all of the past research mentioned so far concerns trying to do predict which *documents/sources* are cited, given the original report document.

To the best of our knowledge, using the report documents at large – regardless of if one assumes knowledge of the contained *citances* – to predict a finer-grain relevance/evidence within the cited documents is a new area of research. That is, instead of predicting the source documents that are cited, the source documents are provided, and the task is to predict the most relevant pieces of text which justify the source being cited.

Using citances, in general, is not entirely novel though: White [11] provides an overview of citation research and states that using citances and their recurring themes of textual content is one of the three main sub-areas of citation research. In this direction, Nakov et. al. [9] leveraged the fact that within the BioMed community, citations are highly plentiful and thus a corpus of bio-medical documents may contain a wealth of citances which all reference the same source document. Therefore, one can use these collections of citances to help construct comparable corpora or new summaries (with respect to the sources).

3 Methods

3.1 Task 1a: Evidence Linking

This sub-task, as mentioned, is concerned with building a model to automatically predict the most relevant pieces of text within a cited/source document. Specifically, for each report's citance, from the provided source document we may return a sentence fragment, full sentence, or up to five contiguous sentences – as these were the specifics that human annotators provided during the creation of truth/gold data, which we explain in Section 4.1.

For all of our Task 1a models, we only considered predicting the most relevant *sentences* within the cited documents – not sentence fragments – which, per our corpus, is a reasonable choice to make, for the human annotators did not often pick *sentence fragments* as being the gold, relevant pieces of content. Further, given a report document and a contained citance, we ranked each candidate sentence within the source while only using the text from the corresponding report document's citance. That is, our models did not use the full report document's text or other reports' citances which cite the same source.

Baselines:

- **Jaccard:** our first baseline model, for a given citance C, ranked each candidate source sentence S according to its Jaccard Similarity with the citance, as shown in Equation 1.

- **Vanilla Sum:** second, we consider just the numerator of *Jaccard*. That is, we simply rank sentences based on the number of shared words they have with the given citance.

- **Longest Substring Match:** third, each candidate source sentence is ranked according to its longest number of consecutive words it precisely shares with the report's citance. This is similar to the other two baselines but with the motivation that longer sequences are more indicative of containing contextually relevant information.

$$J(C,S) = \frac{C \cap S}{C \cup S} \tag{1}$$

Weighted Jaccard:

We experimented with two ways of weighting the words so as to alter the Jaccard Baseline. Our assumption was that not all words within a citance and candidate source sentence are equal. Therefore, we ran the Latent Dirichlet Allocation (LDA) [1] topic model over our entire corpus in order to construct 50 topics, each of which has its own most topical words. Our two models are:

- **Topically-Weighted per Report Document:** Akin to Jaccard Similarity, only the words found within the report's citance and candidate source sentence will be used; however, instead of each word effectively having a weight of 1, each word is weighted per Equation 2.

- **Topically-Weighted per Source Document:** Identical to the above, but we use the source document's topic distribution to define the importance of the word – as shown in Equation 3.

$$\text{word weight } w = \sum_z P(w|z)P(z|report) \qquad (2)$$

$$\text{word weight } w = \sum_z P(w|z)P(z|source) \qquad (3)$$

Topic Similarity:

Additionally, we experimented with modelling the importance of each sentence by how topically similar it is to the corresponding citance. Namely, we again ran LDA over our corpus (50 topics), then each sentence (i.e., the report citance and candidate source sentence) was assigned a topic distrubtion per Equation 4.

$$P(z|sentence) = \prod_{w \in sent} P(z|w) \approx \prod_{w \in sent} \frac{P(w|z)P(z)}{\sum_{z'} P(w|z')P(z')} \qquad (4)$$

Our two models differed only in their metric for comparing topic distrubtion similarity:

- **KL-Divergence**

- **Cosine Similarity**

Discriminative Approach:

We identified several features which we thought to be useful in determining if a candidate source sentence is pertinent to a given citance:

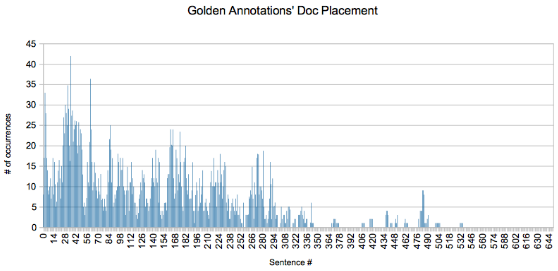

Figure 1: Document placement of golden, human annotated sentences

- bag-of-word (BoW) representation encoding both citance and candidate source sentence (1 if word is present in both; 0 otherwise). We limited our vector's vocabulary to the 6,000 highest ranking words per the aforementioned LDA topic model.

- highest rank position of an adjacent candidate sentence (per Vanilla Sum baseline)

- candidate sentence's length

- candidate sentence's placement within the source document (0 represents 1st sentence, $N-1$ represents last sentence)

The last feature was chosen after looking at our data and noticing there is a tendency for the golden, human annotated sentences to be ones closer to the top of the source documents, as shown in Figure 1.

Specifically, we used these features and performed logistic regression with squared L2 regularization. Our training and testing details are listed in Section 4.1.

3.2 Task 1b: Discourse Prediction

This sub-task concerned building a model to automatically categorize cited text spans in the referenced paper into one of the following 5 facets: hypothesis citation, method citation, results citation, implication citation and discussion citation.

For our model, we took an ensemble approach that exploited a bag-of-word (BoW) representation encoding both citance and (predicted) reference text features (more on this below). More specifically, this ensemble included the following models: multinomial logistic regression with squared L2 regularization; linear kernel SVM; random forest (with 20 base learners); k-nearest neighbors and multinomial Naive Bayes. We tuned all model hyperparameters – including the regularization penalty for logistic regression and SVM, the k parameter for k-nearest neighbors and the smoothing parameter for Naive Bayes – via grid-search over the training data to maximize predictive accuracy.

Many sentences were associated with more than one label (i.e., annotators disagreed with one another). To account for this, we simply included multiple copies of each sentence: one for each label in the corpus. Thus if a sentence was labeled once as Hypothesis Citation and twice as Method Citation, we would include three copies in our training dataset with these labels. As a result, the disagreement is implicitly accounted for during training, as mispredicting sentences uniformly agreed upon would contribute more to the empirical error than 'mislabeling' a sentence about which annotators disagreed.

Concerning features, we included all uni- and bigrams that occurred in at least 5 citance texts. We did not perform stop-wording. We used term frequency inverse-document frequency representation (tf-idf).

Intuitively, we anticipated that both the text in the citance and the relevant text in the cited document would be predictive of the facet. Therefore, we incorporated features extracted from both texts, where these are kept separate: i.e., features extracted from citances are treated independently from those extracted from the reference texts. At train time, this is straight-forward because the relevant reference texts are known.

At test time, we substituted features from the 'true' reference texts *predicted* with features extracted from texts predicted by our model (from task 1a) as being most likely to correspond to the citance under consideration. Using features from the 'true' reference texts improved performance substantially. Substituting the predicted texts in still seemed to improve performance, although by considerably less. One may hypothesize that improving our model for task 1a would likely also improve our model for 1b.

4 Experiments

4.1 Data

As mentioned, NIST sponsored the TAC 2014 Biomedical Summarization Workshop, which motivated this work. As part of their effort, they provided a rich corpus of 20 distinct source documents. Each of these source documents has an associated 10 distinct report documents which cite the given source (and the citance is specified). For each of these 200 pairs (20 sources, each linked by 10 distrinct reports), 4 human annotators manually read the entire source document and selected the most relevant sentence fragment or sentence(s) – limited to 5 contiguous sentences. Naturally, there was some disagreement amongst annotators. On average, most source documents had roughly 200 sentences, with one outlier containing 624.

Additionally, the workshop also provided 30 more source documents. Again, each of these had an associated 10 report documents, and humans annotated all citances. However, this set served as the test set, so the annotations were not provided and shall be used for shared-task's evaluation.

In lieu of having the answers to the test set, we evaluated our models by dividing the training set into 2 groups: we randomly selected 16 of the sources to serve as training and the remaining 4 for test. For all of our experiments, this training/testing split remained constant.

4.2 Entity Linking

Each of our models, for a given citance, returns a ranked list of the source's sentences. Each sentence can be represented by the character offset within the source document (e.g., the first sentence could be bytes/characters 100 - 120 if the first 99 bytes of the document were header information). Since the human annotators were allowed to select sentence fragments and not just complete sentences, it makes sense to evaluate based on the number of overlapping bytes between our models' predictions and the humans', a la typical recall and precision metrics. Due to the natural disagreement that occurs between human annotators, the scoring metric set forth by the workshop (which we also used for our experiments) is a weighted F1 metric. For a given citance, we define $WeightedRecall(S|M)$

$$WeightedRecall\ (S|M = \{G_1, G_2, \ldots G_m\}) \stackrel{\text{def}}{=} \frac{|S \cap G_1| + |S \cap G_2| + \cdots + |S \cap G_m|}{|G_1| + |G_2| + \cdots + |G_m|}$$

$$WeightedPrecision\ (S|M = \{G_1, G_2, \ldots G_m\}) \stackrel{\text{def}}{=} \frac{|S \cap G_1| + |S \cap G_2| + \cdots + |S \cap G_m|}{m \times |S|}$$

Figure 2: Weighted Recall and Weighted Precision

and $WeightedPrecision(S|M)$ for a model returning a set of indexed bytes S, with respect to a set of M annotations from m humans, containing indexed bytes $G_1, G_2, ..., G_m$ according to Figure 2. Then, we define weighted F1 ($wF1$) per Equation 5.

$$wF1 = 2 * \frac{WeightedRecall * WeightedPrec}{WeightedRecall + WeightedPrec} \quad (5)$$

Having trained on the 16 randomly selected sources and tested on the remaining 4, we show our average performance across all citances in Figure 3. The x-axis represents the number of sentences we return for a given source; hence, performance for the first few indices are most important, especially since for each source humans typically selected roughly 3 sentences as being relevant.

As shown, the Logistic Regression Discriminate approach performs the best; however, the Vanilla-Sum baseline performs very well and was difficult to outperform. The "1 perform anno + random" model represents the highest performance one could reasonably expect, for we randomly selected a human annotator's answers and measured performance against the other 3 humans. Since only 3-5 sentences were typically selected as truth data, for the remaining number of sentences returned, we selected randomly. The topic-modelling-based similarity models perform poorly, suggesting that the actual distinct words found within sentences and citances matter more than simply containing topically related words.

5 Discussion

Having looked at the data, it is clear that citances correspond to at least three different *types* of relevant source sentences:

1. **Keyword-based:**
   citance: *The general impression that has emerged is that transformation of human cells by Ras requires the inactivation of both the pRb and p53 pathways, typically achieved by introducing DNA tumor virus oncoproteins such as SV40 large tumor antigen or human papillomavirus E6 and E7 proteins.*

   1 human's annotated source sentence: *Several viral oncoproteins also cooperate with ras, including SV40 T-antigen, adenovirus E1A, human papillomavirus E7, and HTLV-1 Tax (70 and 57). When expressed alone, most of these cooperating alterations facilitate the establishment of primary cells into immortal cell lines*

2. **Paraphrasing & Lexicalized Reordering:**
   citance: *Loss of TET2 is believed to cause an aberrant methylation of promoter regions in AML*

   1 human's annotated source sentence: *TET2 loss of function would be anticipated to result in hypermethylation, and the data reported here support this scenario*

3. **Summarization & Topics:**
   citance: *Recent analyses of multiple different cancers have identified gene expression differences between tumors with similar histologic characteristics yet heterogeneous clinical behavior*

   1 human's annotated source sentence: *As shown, distinct groups of genes distinguish cases defined by E2A-PBX1, MLL, T-ALL, hyperdiploid >50, BCR-ABL, the novel subgroup, and TEL-AML1. In addition to these specific subgroups, 65 cases (20% of the total) were identified that did not cluster into any of the leukemia subtypes*

We used topic modelling in attempt to capture keyword-based source sentences and hopefully some of the "summarization & topics" sentences. However, especially for the latter, it appears one would need a model that truly incorporates a richer model – the example citance mentioned "multiple different cancers"; however the humans' golden sentences never mention these words. Instead, the rel-

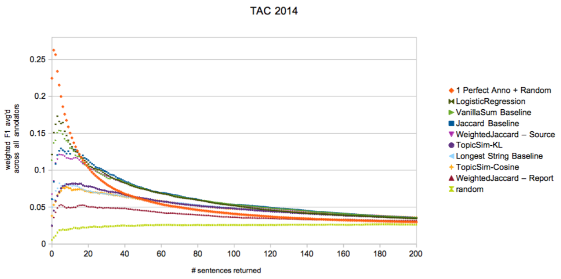

Figure 3: Performance of all models on Task 1a

evant sentences actually list specific cancers, suggesting that the citance was truly a summarized fact which introduced new words not found in the individual sentences. Without building more complex models to handle these various cases, it seems reasonable that our VanillaSum model could do so well, as it is a simple approach that generally captures common relevance.

6 Conclusion

We introduced several basic models in attempt to find relevant sentences (aka link evidence) to a corresponding, provided citance. Per our preliminary results, our descriminative logistic regression approach offered the best performance, while VanillaSum (simply the number of shared words between the citance and candidate sentence) yielded competitive results.

For the discourse prediction (facet identification) task, our ensemble classifier approach offered the best results, yielding 52% accuracy on our held-out evaluation data. Had we performed perfectly at the evidence linking task, then using the predicted, relevant source sentences as features can potentially increase performance to as much as 62%. However, with our current performance at evidence linking, we were able to increase performance to just 54% accuracy.

It is clear that richer models might be necessary for the task of evidence linking, as citances often refer to many different types of relevant content. Further, it might be beneficial to use our discourse facet predictions to help the task of evidence linking – if we are fairly certain that a citance corresponds to a "discussion", then our predicted sentence should likely be from the "Discussion" section of the document.

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] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220–5227, 2004.

[3] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430. ACM, 2010.

[4] David Cohn Thomas Hofmann. The missing link-a probabilistic model of document content and hypertext connectivity. In *Proceedings of the 2000 Conference on Advances in Neural Information Processing Systems. The MIT Press*, pages 430–436, 2001.

[5] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[6] Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C Lee Giles, and Lior Rokach. Recommending citations: translating papers into references. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1910–1914. ACM, 2012.

[7] Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. Utilizing context in generative bayesian models for linked corpus. In *AAAI*, volume 10, page 1, 2010.

[8] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link lda: joint models of topic and author community. In *proceedings of the 26th annual international conference on machine learning*, pages 665–672. ACM, 2009.

[9] Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. Citances: Citation sentences for se-

mantic analysis of bioscience text. In *Proceedings of the SIGIR04 workshop on Search and Discovery in Bioinformatics*, pages 81–88, 2004.

[10] Ramesh M Nallapati, Amr Ahmed, Eric P Xing, and William W Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550. ACM, 2008.

[11] Howard D White. Citation analysis and discourse analysis revisited. *Applied linguistics*, 25(1):89–116, 2004.