A Baseline Approach to the RTE5 Search Pilot

Andrew MacKinlay and Timothy Baldwin
NICTA VRL
University of Melbourne
Melbourne VIC 3010, Australia
amack@csse.unimelb.edu.au, tb@ldwin.net

Abstract

1 Introduction

The 2009 Recognising Textual Entailment (RTE5) Search Pilot was a shared task in which the aim was to find sentences in a collection of documents which logically entailed particular "hypothesis" sentences. The data set was a collection of newswire documents from various sources, arranged into particular topics, and segmented into sentences, along with a set of hypotheses for each topic. Each sentence in the set of documents associated with a given topic was potentially involved in an entailment relationship with each hypothesis for the topic; the task is to identify for which sentence/hypothesis pairs this was actually the case. In the gold-standard annotations of text–hypothesis relationships, only a small proportion of the sentence of the topic were taken as entailing a particular hypothesis. The principle differences over previous RTE tasks were that each hypothesis was associated with a set of documents rather than a unique "text" (conventionally a single sentence), and the data was heavily skewed towards negative entailment.

We present here our approach to this task which, despite using only simple bag-of-words based techniques — primarily variants on cosine similarity between the text and hypothesis — achieves competitive results over the development and test data.

The RTE task, and in particular the Search Pilot, is of interest to us because it is a well-defined task with wide reach that is somewhat related to a subject that is a more primary research concern: identifying instances of **semantic duplication** in document collections. In online document collections, users often wish to identify whether incoming documents have close semantic matches in the existing collection, i.e. the specific content or topic of the incoming document matches with that of a previously-published document. We consider semantic duplication to occur over a document pair $\{A, B\}$ in one of two forms: (1) **subsumption**, where A properly subsumes B (i.e. $B \subset A$); and (2) **synonymy**, where A and B have identical semantic content (i.e. $A \equiv B$), that is the two documents subsume one another. Subsumption here has obvious parallels to entailment in RTE, although the definition of subsumption is somewhat more flexible and specific to the task rather than derived in principle from inferential logic — in particular, we do not wish to be restricted to propositional statements. The domains we

intend to apply this work to are also somewhat different — the focus is not on newswire but rather on several distinct technical domains.

The original motivating context for this research is web user forums, and specifically technical troubleshooting forums. In technical domains such as Linux operation/configuration, web user forums are the primary means to service users' needs, through users posting complex questions that cannot be phrased easily in a short post, and other users/volunteer "gurus" answering them asynchronously. Within the substantial body of information in any popular forum, there is a good chance that a given question will have been answered in the past. If there were the facility in a forum to automatically redirect previously-answered questions to relevant historical data, the person with the problem would receive a more immediate answer, and guru time could be used more effectively in answering novel questions, rather than providing links back to old threads where the same questions were asked and answered.

While this task of finding questions that subsume the novel question is clearly a distinct task to RTE, and the exact same approaches will likely not work for both, in each case we are attempting to locate in a large collection (either forum posts or newswire sentences) the small subset of documents that have salient semantic commonalities with a particular target (an incoming forum question or a hypothesis). It would therefore be unsurprising if some techniques were applicable to both tasks.

2 Methodology

2.1 Basic Technique and Baseline

This work is intended as a preliminary investigation of the task, evaluating how far we can get with simple techniques such as bag-of-words, as well as off-the-shelf tools. There are a number of reasons we would want to use relatively standard and naive techniques based on bag-of-words for this task.

The Search Pilot task is more similar in nature to an information retrieval (IR) task than the established RTE task. Where in the main task we would compare only those relatively small number of text/hypothesis pairs that are licensed by the dataset, in the Search Pilot, each hypothesis could potentially be entailed by any of the text sentences for the topic that the hypothesis applies to. Thus, rather than simply evaluating whether each of a relatively small number of pairings is an instance of entailment, every possible within-topic pairing of each hypothesis with each sentence in the texts (19294 such pairings in the development set) is a potential instance of entailment. We are thus trying to pick out which small subset of documents (in this case sentences) match a certain set of criteria (which is entailment here); this has clear parallels to a search task in IR of finding subsets of documents that are relevant to a query. This is not to imply that the search task is an IR task — we are looking for a more clearly defined notion of entailment rather than a fuzzy notion of relevance, and the scale of the collection is not what would usually be seen in IR — but it does indicate that a naive IR-type approach might be an interesting starting point.

Intuitively it seems likely we would expect a relatively high level of lexical overlap between a text and the corresponding entailed hypothesis. For entailment to occur, it seems probable the text will at least mention some terms that also appear in the hypothesis. For example, the following text/hypothesis pair is annotated an instance of entailment in the development set, and there is a large number of overlapping terms between the two:

H: *The Airbus A380 flew its maiden test flight.*

T: *Airbus A380 takes off on second test flight*

An established solid performer in the IR domain for measuring lexical overlap is cosine similarity. Thus it seems reasonable in the first pass to use cosine similarity with bag-of-words features over the text and hypothesis, as this will give us a single numerical measure of the similarity of the two sentences. We also follow the common practice of weighting the terms by their inverse document frequency, as this prevents relatively frequent terms from unfairly dominating the similarity calculations. Specifically, for the IDF weighting for term t, we use the well-known calculation:

$$\log \frac{N}{d_t}$$

where N is the number of documents and d_t is the number of documents in which t appears.

While it is possible to assign a similarity score between a text and hypothesis with our cosine similarity formulation, this alone does not enable us to make entailment predictions. The most obvious means of using that score is to determine a threshold, above which we mark a T–H pair as entailed. In our case, we determine the threshold which would maximise the F-score over positive instances (i.e. entailment) in the training data, and apply that threshold to the test data. We do not separate the topics from each other in calculating the threshold, but selecting the threshold on the basis of macro-averaged optimal thresholds from the different topics would be an interesting alternative to investigate. We performed experiments both using the standard IR practice of stemming the terms (in our case using Snowball[1]), as well as explicitly lemmatising them (falling back to stemming if this failed).

2.2 Ignoring Irrelevant Sentence Portions

We also experimented with an extension of basic cosine similarity, based on the observation that texts which entail a hypothesis frequently contain a large amount of irrelevant information. For example:

H: *The Airbus A380 flew its maiden test flight.*

T: *French President Jacques Chirac immediately hailed the "total success of the first test flight of the Airbus A380".*

[1] `http://snowball.tartarus.org`

In the cosine similarity metric, the similarity score for pairs such as these is lower than we would like, since the additional terms that appear in the text and not the hypothesis lower the similarity score of the two documents, due to the normalisation of the bag-of-words vectors (rarer terms such as *Chirac*, with their higher IDF, have even more of an impact here). This is often a desirable property for approximating relevance in IR, but here it is not ideal. A simple ad-hoc variant of cosine similarity can avoid this problem: we remove all terms from the vector for the text that don't also appear in the hypothesis.[2] In this way, we are only examining terms that we expect to impact on the entailment determination, and accept that entailing texts can easily contain irrelevant information. This is not a particularly principled approach, but as we shall see below, it is relatively successful empirically.

2.3 Adding Synonymy

Unsurprisingly, in some cases words from the hypothesis don't appear in the text, but synonyms of those terms do:

H: *The ice is **melting** in the Arctic.*

T: *That study, commissioned by the United States and seven other nations, found permafrost there to be **thawing** and glaciers and sea ice to be retreating markedly, raising new concerns about global warming and its impact.*

Ideally we would like to match on near-synonyms such as these[3] and augment the corresponding bag-of-words feature vectors if only a synonym is found. The most readily available off-the-shelf system for this is WordNet (Miller 1995). The algorithm we use is to lemmatise each term, and fetch all lemmas for all synsets of that lemma (not making use of the WordNet hierarchy). For each of those lemmas, if we have IDF counts for the term in question, we augment the vector of term counts for that lemma, possibly discounting the weight by some factor close to 1. We also experimented with expanding the set of terms with derivationally-related forms (according to WordNet) in an almost identical fashion, to capture similarities such as that between *mine* and *miner* in the following:

H: ***Mine** accidents cause deaths in China.*

T: *More than 6,000 **miners** died in accidents in China last year, according to previously released government figures.*

[2] In fact, a bug discovered after submission meant that each term in the text was assigned the term count of the corresponding term in the hypothesis. The main practical outcome of this would be setting the term counts to 1, since there are almost no significant repeated terms in the hypotheses. In any case, we would not expect this to make a substantial difference, since there are relatively few repeated terms in the texts, as they are single sentences. Empirically, it also made little difference (a change in the precision, recall and F-score in the 3rd significant figures) but we report results over the version with this minor bug since that is what was submitted.

[3] The fact that it is *permafrost* rather than *ice* which is thawing in this example is irrelevant in a bag-of-words model, particularly since *ice* is mentioned later in the sentence.

3 Results

We present our results over the test and development data in Table 1. For development results, we treated each topic as a fold for leave-one-out cross-validation.

Unsurprisingly, plain cosine similarity with IDF weighting is a solid performer, both over the test and the development sets. Its F-score of 0.353 over the test set is substantially above the reported median for all submitted runs of 0.301, although since this figure includes all runs from each team, it is not necessarily competitive with each team's best run. There is only a slight reduction in performance over the test data, which is well within what we would expect.

The substantial performance boost achievable by making the minor modification to the cosine similarity method, in pruning the text feature vector, is interesting. It seems that our intuition about irrelevant portions of sentences making the cosine similarity metric less accurate were correct, and simply removing the superficially irrelevant terms makes a large difference to this – again approximately maintaining this performance over the development data.

Augmenting the counts with WordNet synonyms was surprisingly unproductive. Adding immediate synonyms (Syn) caused a substantial drop in performance over the development data compared to the corresponding run without synonyms, and adding derivationally-related forms (DRFs) caused a slight further decrease. At first glance, it may be surprising that the recall decreased, when we might naively expect an increase if we expand the comparison vectors, but since we are optimising to maximise F-score over the training data, it is not necessarily the case that the recall should increase — to bring the precision up to an acceptable level, the calculated similarity threshold may be much higher. It is also not clear why there is such a difference in recall over the test set. Presumably the synonymy expansion was more productive there due to the presence of certain terms, although the corresponding drop in precision more than offsets any gains made here.

4 Conclusion

We have developed a system based on variants of standard document comparison techniques which, with minimal implementation effort, produce reasonable performance over the supplied datasets. Apart from the obvious advantage of simplicity of implementation, techniques based on bag-of-words, or more specifically those which can be implemented using inverted indexes, can also be implemented relatively efficiently. Any real world system which uses more sophisticated techniques on a large scale document collection will probably need to find a way to avoid having to run an expensive comparison algorithm between the hypothesis and every document in the collection. Techniques based on inverted indexes are a plausible candidate to use for pre-filtering, as long as they can be shown to achieve reasonable performance — in particular high recall while still returning only a small percentage of documents. This is not to imply that our

Metric	WordNet	Terms	Devel.			Test			
			P	R	F	P	R	F	Run
Cosine	—	Stem	.339	**.437**	.382	.293	.445	.353	2
Cosine	—	Lemma	.344	.407	.373	—	—	—	—
SubvCos	—	Stem	**.460**	.421	**.439**	**.429**	.380	**.403**	1
SubvCos	—	Lemma	.439	.400	.419	—	—	—	—
Cosine	Syn	Lemma	.244	.319	.276	.194	**.486**	.278	3
Cosine	Syn, DRF	Lemma	.215	.335	.262	—	—	—	—

Table 1: Results over the development (using leave-one-out cross-validation by topic) and test data (trained on the development set) matching documents above a learned similarity threshold (SubvCos = cosine similarity over the pruned feature vector for the text; Syn = synonym-based WordNet expansion; DRF = WordNet expansion using derivationally-related terms; Run = the ID of the officially-submitted run)

system would be appropriate for this task in its current form, but it does suggest the plausibility of using something similar.

Obviously it would be valuable to apply a more sophisticated algorithm to the task. As mentioned in the introduction, a primary focus of our research interest is on the related task of detecting instance of semantic duplication focussed on Linux technical forum questions, among other tasks. Preliminary work indicates that similar bag-of-words produce similarly robust performance, but it is likely that more sophisticated techniques will be required. One avenue of research is using the semantic output of the English Resource Grammar (Copestake & Flickinger 2000) to elucidate more complex semantic relationships, and it possible that some of these techniques will also be applicable to the RTE task.

References

Copestake, Ann & Dan Flickinger: 2000, 'An open source grammar development environment and broad-coverage English grammar using HPSG', in *International Conference on Language Resources and Evaluation*.

Miller, G.A.: 1995, 'WordNet: a lexical database for English', *Communications of the ACM*, **38**(11): 39–41.