Summarizing with Encyclopedic Knowledge

Vivi Nastase
EML Research gGmbH
Heidelberg, Germany
http://www.eml-r.org/nlp

David Milne
Waikato University
Hamilton, New Zealand
http://www.cs.waikato.ac.nz/d̃nk2/

Katja Filippova
EML Research gGmbH
Heidelberg, Germany
http://www.eml-r.org/nlp

Abstract

This paper presents a topic-driven multi-document summarization approach that relies on linking documents to Wikipedia. Wikipedia provides structural support to retrieve relevant concepts from the documents to be summarized, and quantify the strength of the relations between them, thus expanding the topic. We identify concepts in the documents, and assign them scores that describe their relevance to the topic, their significance in general, and a machine-learned confidence that they should appear in the summary. Sentences are ranked according to the scores of the concepts within them and how much new information they provide. The best are extracted and compressed to form the summary. The system is trained and developed using the DUC 2005 and 2006 data. It was tested on the DUC 2007 data before deploying it on the update summarization task of TAC 2009. It performs 5th (compared to 30 peers) in DUC 2007, and 21st (of 52 peers) on the TAC 2009 update task.

1 Introduction

In this edition of the summarization competition we have experimented with a novel approach to topic/query expansion. Because queries describing a user's information needs are rather short, expanding them is crucial to retrieving the most relevant information from a set of documents. Previous approaches have used WordNet, additional sets of relevant and irrelevant documents for a topic, or clusters of closely related terms from the documents that are being summarized. In this paper we describe a novel approach, which links documents to Wikipedia, and thus allows for an expansion, and ultimately inter-connection, of the query and document terms within this large semantic space. Because it takes into account both the quantity and quality of concepts within a given sentence relative to the topic, the system is biased towards longer sentences. To counter this – given the fact that the task is to generate 100 word summaries – a sentence compression module produces more compact sentences by pruning unnecessary phrases.

Multi-document summarization involves expressing concisely the most relevant information contained within a set of documents, relative to a particular topic or information need. Extracting and compressing information this way is becoming increasingly desirable, given the ever accelerating growth of data we must cope with on a daily basis.

If we analyze our own way of digesting documents, it is obvious that we often draw on knowledge that is not explicitly given in the text. Imagine, for example, that one is curious about the cast and crew of the hit comedy ”Seinfeld”, and how they have coped since the series was cancelled. A recent news article declares that ”Michael Richards is bringing his gift for slapstick to a new NBC sitcom.” Background encyclopedic knowledge is essential to understand this sentence, and to judge its relevance to the information need. Unless one understands that *Michael Richards* is the actor who played *Kramer* – an iconic Seinfeld character – the sentence is meaningless.

The same knowledge should be made available to automatic summarization systems. Lexical semantic knowledge – wordnets, dictionaries and thesauri – has been used extensively in summarization research. Large-scale, open-domain, machine-readable encyclopedic knowledge, however, has only recently become available and remains largely untapped for summarization. The online encyclopedia Wikipedia is a particularly promising source of world knowledge, and has been applied to a host of natural language processing and knowledge management tasks.

The system described here makes extensive use of Wikipedia. The resource provides a large inventory of articles, which we use to represent the concepts discussed within the documents to be summarized. It also provides a large and highly connected network of semantic relations, which we use to quantify how document concepts relate to topic descriptions and the manual summaries that satisfy them. This allows us to learn which concepts are relevant for expanding upon and responding to the original information need.

The remainder of the paper is structured as follows: an overview of related work is presented in Section 2. The following three sections correspond to the main steps of our own approach: identifying the Wikipedia concepts discussed in documents (Section 3), learning to identify concepts that should–and should not–be discussed within summaries (Section 4), and forming the final summary (Section 5). Section 6 evaluates the knowledge-based summarizer, which ranks high compared to other participating systems. The remaining section discusses the strengths and limitations of our approach, and points out some directions for future work.

2 Related Work

The main venues for document summarization research are the Document Understanding Conference (DUC) and its successor (since 2008) the Text Analysis Conference (TAC)[1]. These NIST organized events provide a forum for the comparison of systems across various summarization tasks. Summarization however is not a novel task – interest in extracting the gist of documents dates back to the 1950s and 1960s (Luhn, 1958; Rath et al., 1961; Edmundson, 1969).

Between 2005 and 2007 one of the DUC tasks was topic-driven multi-document summarization. The organizers provided a number of topics, each consisting of a title and a set of sentences and questions that describe an information request. Each topic has an associated set of approximately 25 documents, from which a 250 word summary should be created. The automatically-generated summaries are evaluated against multiple (typically four per topic) manually-defined summaries. The systems are manually and automatically evaluated, and scored for linguistic quality – or readability – and responsiveness – the extent to which they answer the information need.

Figure 1 shows one of the topics from DUC 2007, where participants are asked to summarize what happened when the hit comedy series *Seinfeld* finally came to a close.

Since 2008, the summarization task has changed slightly – it is still a topic-driven multi-document summarization task, but it requires shorter summaries (100 words) in several (2) steps: A). an initial summary of 10 documents on the given topic, and B). an update summary (also from 10 documents) on the given topic, but with novel information (compared to the summary generated in step A).

```
<topic>
<num> D0739I < /num>
<title> after "Seinfeld" < /title>
<narr>
What became of the cast and others related to the "Seinfeld" TV series after it ended? What actions were taken
by others in response to the show's closing? < /narr>
<docs>
...
< /docs>
< /topic>
```

Figure 1: Sample topic from DUC 2007

Over time we have seen a wide variety of approaches to this problem, ranging from knowledge poor to knowledge rich; from simple sentence ranking to complex graph-based algorithms or machine learning. Most systems perform extractive summarization, with some attempts to trim or compress sentences by deleting phrases such as relative clauses or parentheticals[2]. The topic descriptions are fairly short, and offer few clues to help decide which sentences are important. A central part of the task is to expand upon these clues.

Statistical and grammatical analysis of corpora offer one option. Lin and Hovy (2000) build "topic signatures" – expansions of topics with related terms – based on sets of documents marked as relevant or not with respect to the topic. This approach is extended by Harabagiu (2004), who enhances the topic representations with pairs of grammatically related words.

Lexical resources, particularly WordNet (Fellbaum, 1998), are also useful. Ye et al. (2005) match words or sequences of consecutive words within candidate sentences to the relevant WordNet entries. For two sentences s_i and s_j, they collect the set of concept pairs $(c_{i,x}, c_{j,y})$ whose similarity scores (computed using a Lesk measure based on their WordNet definitions) are greater than a threshold – thus disambiguating words to their closest senses. The similarity between sentences is then computed as the weighted sum of the strength of each concept pair in this set. Based on the sentence similarities, Ye et al. compute several scores – representative power, similarity with the topic, etc. – and choose sentences for the summary based on a modified Maximum Marginal Relevance (MMR).

Recently, Wikipedia has emerged as a useful resource for summarization. It represents a very large inventory of concepts (2,000,000+ articles for English), most of which are named entities – people, places, events – which lexical resources do not aim

[1] http://duc.nist.gov, http://www.nist.giv/tac

[2] http://www-nlpir.nist.gov/projects/duc/pubs.html

to cover. Additionally, the concepts are augmented with extensive descriptive text that is not available in such resources, and a large amount of semi-structured knowledge (e.g. categories and hyperlinks) that other corpora lack.

Svore et al. (2008) link news articles to Wikipedia and add to position and word frequency-based sentence scores two features that capture the importance of a sentence relative to Wikipedia entities. These scores boost the importance of sentences that contain entities frequently mentioned in the news agency's documents. The summarization system is a supervised method that learns how to choose the three best sentences as a summary, using manually written three sentence summaries for training.

Biadsy et al. (2008) develop an unsupervised multi-document extractive summarization system to produce biographies. Sentences from Wikipedia biography pages provide instances for the positive class, while negative instances are extracted from the general news corpus TDT4. Data is preprocessed with a NE tagger and coreference resolution system, and sentences are represented through n-gram and POS features.

Unfortunately neither of the Wikipedia-based approaches described above fall within the DUC framework, making them difficult to compare against. To our knowledge, only Nastase (2008) applies Wikipedia to DUC tasks, and its use of Wikipedia is quite peripheral. This system builds a graph from the documents to be summarized, where nodes are open-class words, and edges are the grammatical dependencies between them. Spreading activation (starting from a cluster of nodes that are related to the topic) is used to weight edges in the graph, and PageRank is used to determine the highest ranking nodes. These nodes are used to score the sentences, and the best ones are selected to form the summaries. Wikipedia's only contribution is to identify the related entities used to initialize the spreading activation process. The system described in the following Sections makes much more direct and extensive use of Wikipedia.

3 Cross-referencing documents with Wikipedia

The first step in summarizing documents is to detect the Wikipedia concepts discussed within them. To achieve this, we draw on the work of (Milne and Witten, 2008b). This section provides a brief overview of the approach.

3.1 Identifying concept terms

Wikipedia contains millions of links, which provide an extensive vocabulary of terms – link anchors – and their target articles. We map document terms to Wikipedia articles – which are functionally the same as concepts – by gathering all n-grams below a certain length and consulting this link vocabulary. Stemming and lemmatization are not required, because synonymy and other surface-form variants have been encoded manually in link anchors: for example, both Wikipedians and journalists have difficulty spelling the name *Seinfeld*, which is also referred to in the link structure as *Seinfeild*, *Sienfield*, and, of course, *the show about nothing*.

One complication for this approach is that the link vocabulary is, if anything, *too* comprehensive. It even covers stopwords such as *and*, *or* and *the*. Mihalcea and Csomai (2007)'s keyphraseness (or link probability) feature is used to discard such unhelpful terms. For each candidate phrase p, the probability of being a concept is:

$$kp(p) = \frac{f_a(p)}{f_t(p)}, \tag{1}$$

where $f_a(p)$ is the number of Wikipedia articles in which it is used as an anchor, and $f_t(p)$ is the number of articles in which it appears in any form. Phrases with low probabilities are discarded.

3.2 Resolving ambiguous terms

Terms and phrases can be ambiguous: *Seinfeld* could refer to the show, to its namesake *Jerry Seinfeld*, to the character that he plays within the show, or to the completely unrelated musician *Evan Seinfeld*. This ambiguity is reflected in the link structure, so that *Seinfeld* links to different destinations depending on the context in which it is found. We use a machine-learned approach to choose the correct destination because Wikipedia provides millions of ground-truth examples to learn from: every link in every Wikipedia article has been manually disambiguated.

The approach is described in detail in (Milne and Witten, 2008b). Briefly, it balances two main features for each possible sense of an ambiguous term: commonness (i.e prior probability) and relatedness to context. The commonness of a sense is defined by the number of times it is used as a destination in Wikipedia. For example, almost all of *Seinfeld* links refer to the television show, while less than 1% link to the actor or the musician. The relatedness of a sense is its average semantic relatedness (discussed in Section 3.3) to all of the concepts that can be mined from unambigous terms in the surrounding text. For training, these are obtained from the Wikipedia article in which the link was found. When testing – in our case, summarizing documents – we mine context terms from the description of the topic and its associated documents all at once, to provide maximum context and ensure that terms are disambiguated consistently across them.

3.3 Measuring relatedness between concepts

Our approach for disambiguating and weighting concepts requires a measure of how strongly two Wikipedia articles relate to each other. We use the WLM measure developed by Milne and Witten (2008a), which measures the semantic similarity of two articles by comparing their incoming and outgoing links. Formally, the relatedness measure between two articles a and b is:

$$rel(a,b) = \frac{\max(\log|A|, \log|B|) - \log|A \cap B|}{|W| - \min(\log|A|, \log|B|)}, \tag{2}$$

where A and B are the sets of all articles that link to a and b respectively, and W is the set of all Wikipedia articles.

4 Predicting pertinent concepts

Having identified the Wikipedia concepts mentioned within topics and documents, the next step is to identify those that are worthy of being included in the summaries. We have developed a machine learning approach to do so, which has much in common with the link detection classifier described in (Milne and Witten, 2008b). The link detector uses Wikipedia articles to learn how to distinguish between topics that should and should not be linked. In a similar fashion, our system learns to identify pertinent concepts from documents and human-generated summaries. Positive examples are the document concepts that are also mentioned in these summaries, while negative ones are those that are not. The features that describe the concepts are presented below.

Relatedness to topic concepts Concepts which relate strongly to the query are more likely to be relevant for the summary. For our Seinfeld task, the characters (*George Costanza*, *Cosmo Kramer*, *Elaine Benes*) and actors (*Jason Alexander*, *Michael Richards*, *Julia Louis-Dreyfus*) should rate more highly than other people (the documents mention *Jim Carrey* and *Ellen DeGeneres*, for example). This is measured by the average and maximum relatedness between each concept and those detected in the task description.

Text overlap with topic Our Wikipedia-derived representation of the topic – essentially a set of concepts – does not capture its full meaning. The Seinfeld topic, for example, is represented only by the concepts *Actor*, *Seinfeld*, and *Television program*. This does not capture our interest in the show's closure. To regain some of this lost information, we compute the overlap between the topic and the definition of the candidate concept (the first sentence of its associated article).

Significance within documents Intuitively, one would expect that concepts which are significant and central to a document are more likely to be found within the corresponding summary than those mentioned in passing. We measure a concept's centrality as its average relatedness to all other concepts that were mined from the document. These scores are gathered across the documents as average and maximum values, to capture concepts that are significant either in just one document, or in all of them.

Concept generality Summaries which discuss specific concepts (Seinfeld's actors or characters, for example) are typically more helpful than those that deal in generalities (e.g. entertainment or acting). The generality of a concept is defined as the minimum depth at which it is located in Wikipedia's category tree. This is calculated beforehand by performing a breadth-first search starting from the Fundamental category that forms the root of Wikipedia's organizational hierarchy.

Link Probability Mihalcea and Csomai's link probability, described in (3.1), is a proven feature for differentiating true concepts from the surrounding prose. Because each concept can be referred to by different surface forms (e.g. *Jerry Seinfeld* and *Seinfeld*) there are multiple link probabilities. These are combined into two separate features: the average and the maximum over the collection of documents.

Location and Spread These features capture information about the location and occurrence of concepts within documents. Frequency and document count are obvious choices, since the more times a concept is mentioned, the more important it is. Another feature is first occurrence, because concepts mentioned in the beginning of documents tend to be more important. The distance between first and last occurrences, or spread, is used to indicate how consistently the document discusses a certain concept. These last two location-based features are normalized by the length of the document, and combined across documents as average and maximum values.

5 Generating the summary

Having learned to predict the significance of each concept encountered in the document collection, the next step is to choose the sentences in which they are found. The selection of sentences must convey the most relevant information with a minimum of redundancy. To this end, a score is computed for each sentence that combines relatedness to the topic, relevance to the summary and other scores, as has been frequently done in summarization since the 1960s (Edmundson, 1969).

5.1 Sentence scoring

A sentence's score combines the concepts' confidence scores, their relatedness to the topic, and the sentence's "aboutness" relative to the topic.

Concept score This score combines the confidence scores $cf(c)$ for the concepts c in the sentence S. Had these confidence scores been used for a binary class problem (should (positive) /should not (negative) appear in the summary), the concepts with a score above a threshold (0.5) will be assigned to the positive class, and those below to the negative. Because classification is not perfect, we use two thresholds – τ_p for the positive concepts, and τ_n for the negative ones. Concepts with a confidence score greater than τ_p will contribute to the positive score of a sentence, while those with a score lower than τ_n contribute to a penalty.

The threshold values are adjusted according to the evaluation of the concept learning phase: should the results show high precision and low recall, we lower the threshold τ_p to allow more concepts to be considered; should the opposite be true, we increase it to minimize the noise. τ_n will be low, to avoid penalizing a sentence for wrongly classified concepts. Because the summarization task is guided by a topic T, we prefer high confidence concepts that are related to the topic. We then consider only concepts c whose relatedness ($r(c, T)$) to the topic is above a certain threshold (τ_r).

The concept score for a sentence S, relative to topic T is:

$$Sc_C(S,T) = \sum_{c \in S, cf(c) \geq \tau_p, r(c,T) \geq \tau_r} cf(c) * r(c,T)$$
$$+ \sum_{c \in S, cf(c) < \tau_n} (cf(c) - 1) * r(c,T)$$

$cf(c) - 1$ reflects the "negativity" of a concept – a confidence score of 0 for an undesirable concept will correspond to a factor of -1.

$r(c, T)$ is the relatedness of concept c to the topic T – in effect the maximum relatedness of c to a concept t in the topic (as defined in Section 3.3) :

$$r(c,T) = max_{t \in T} r(c,t)$$

Topic score This score captures how strongly a sentence relates to the topic. The relatedness of each concept to the topic is its maximum relatedness to one of the topic's concepts. The relatedness of the sentence as a whole is:

$$Sc_T(S,T) = \sum_{c \in S, r(c,T) \geq \tau_t} r(c,T)$$

where τ_t is the threshold that limits the relatedness score, $r(c, T)$ is concept c's maximum relatedness to a concept in the topic T described above.

Discourse feature The relative positions of the concepts indicate the extent to which they are central to the sentence. Consider the following:
Prime Time News had to compete with popular sitcoms like Seinfeld.

It mentions note-worthy concepts like *sitcom* and *Seinfeld*, but is clearly not relevant to the topic. The position of the concepts is important. This is captured by:

$$Sc_D(S,T) = 1 - \frac{min_{t \in T} position(t)}{length(S)}$$

where $position(t)$ is the position of topic word t in sentence S.

Top ranked nodes The sentence scores described above focus exclusively on concepts (or nouns). They emphasize sentences that describe as many relevant concepts as possible, without considering the prose that connects them. Unfortunately concept-rich does not always equate to helpful. Consider following sentence, which is not at all relevant to the Seinfeld topic:
On the first page are Jerry Seinfeld's loopy but legible autograph, a hasty-looking flourish from Michael Richards (Cosmo Kramer) and the almost unreadable signatures of Jason Alexander (George Costanza) and Julia Louis-Dreyfus (Elaine Benes).

To compensate, topics are expanded using the technique described in (Nastase, 2008) and Section 2. The documents are first processed with the Stanford

Parser, with the result in dependency relation format (de Marneffe et al., 2006). A graph is built from this representation: the nodes are open-class words or Wikipedia concepts, and the edges are the grammatical relations between them. As before, spreading activation (starting from a cluster of nodes that are related to the topic) weights the edges in the graph, and PageRank determines the highest ranking nodes. This method emphasizes the verbs, adjectives and adverbs that connect topic and topic-related words and concepts. The score for the sentence is

$$Sc_{T_{PR}}(S,T) = \frac{|T_{PR} \cap S_n|}{|S_n|}$$

where T_{PR} is the set of top-ranked nodes, and S_n is the set of nodes in S.

Sentence salience Sentences that cover more information – that overlap most other sentences – are considered more important. LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004) are methods for ranking sentences based on their relations (similarity/relatedness) with the other sentences in the documents. We plan to incorporate such information in the future. For now our formula is:

$$Sc_{Sal}(S,T) = |\{(c,\mathcal{R},\_)|(c,\mathcal{R},\_) \in \mathcal{R}_s, (c,\mathcal{R},\_) \in \mathcal{R}_{S_i}\}| \\ +|\{(c,\mathcal{R},\_)|(c,\mathcal{R},\_) \in \mathcal{R}_s, (c,\mathcal{R},\_) \in \mathcal{R}_T\}|$$

where $(c, \mathcal{R}, \_)$ is a tuple representing the grammatical dependency relation $\mathcal{R}$ between c and an unspecified concept. We allow such partial matching to find information that completes the topic. The score above captures the matching between a sentence and each of the others in terms of grammatical dependency relations, and the matching between the sentence and the topic.

The final sentence score is a linear combination of the scores presented above:

$$Sc(S,T) = w_C * Sc_C(S,T) + \\ +w_{T_{PR}} * Sc_{T_{PR}}(S,T) + w_D * Sc_D(S,T) \\ +w_T * Sc_T(S,T) + w_{Sal} * Sc_{Sal}(S,T)$$

These weights are determined empirically (see Section 6).

5.2 Forming the summary

Constructing the final summary is simply a matter of concatenating the highest ranking sentences, with one caveat: sentences often overlap, and redundancy should be minimized. Thus building the summary is an iterative process, where the ranked candidate sentences S_c are compared against each sentence S_s that has previously been added to the summary. They are added if they offer new information, and ignored if they do not. The extent of new information is determined based on the lexical overlap after removal of stop words ($o(S_c, S_s)$), normalized by sentence length.

If $max_{S_s \in Summary} o(S_c, S_s) \leq \tau_o$ (where τ_o is an overlap threshold), S_c will be included in the summary. If $\exists S_s \in Summary$ such that $o(S_s, S_c) \geq \tau_o$, sentence S_s in the summary is replaced with S_c, from two (highly) overlapping sentences, the longer one is considered more informative. The value of τ_o is determined during parameter tuning (Section 6). The process is repeated until the desired summary length is reached. The last sentence may be truncated to abide by the word limit (250 for DUC 2005-2007, 100 for TAC 2008-2009).

5.3 Sentence compression

The update task limited the size of the summary to 100 words only, making space even more valuable than in the older task of query-focused summarization where the limit was 250 words. In this situation sentence compression techniques are of a great use as they allow us to fit more important content within the summary's length limit. To eliminate irrelevant information from the top-ranked sentences, we used an improved version of our earlier sentence compression method (Filippova and Strube, 2008). The top-ranked sentences were compressed and then added to the summary till the 100-words limit has been reached.

In a nutshell, our sentence compressor proceeds as follows: sentences are parsed with the Stanford parser (Klein and Manning, 2003); the dependency trees are pruned with corpus statistics so that the resulting tree has the maximum syntactic and relevance score possible under a handful of structural constraints; the nodes of the tree are ordered as in the original sentence. Prior to pruning, dependency trees are transformed so that the dependency representation becomes more semantically motivated. For example, function words are eliminated and some dependencies absent from the input are introduced. The size of the pruned tree – i.e., the exact number of dependency edges left – depends on the size of the input tree and has been estimated from a corpus of handcrafted compressions.

6 Evaluation

This work was developed and evaluated within the DUC and TAC framework. DUC 2005 (50 topics and the associated documents and manually-defined summaries) provided training for the concept prediction classifier. Similar data from DUC 2006 (another 50 topics) was used for development – tuning the thresholds and the weights in sentence scoring computations. Prior to entering the TAC 2009 update summary competition, the system was evaluated on the DUC 2007 (45 topics) and TAC 2008 update summary tasks. In all cases the relevant encyclopedic information was mined from a version of the English Wikipedia released in late July, 2008. The Weka toolkit and a bagged C4.5 decision tree learner (Quinlan, 1993) was used for the learning experiments.

We evaluate the outcome of two processing steps: predicting the concepts that should be included in the summary and scoring the sentences to produce the summaries.

Quality of predicted concepts Figure 2 plots how the concept scoring classifier identifies summary-worthy concepts for 2006 and 2007 topics when trained on 2005 data. F-measure is plotted against a threshold τ, which binarises the confidence scores produced by the classifier into positive and negative classes.

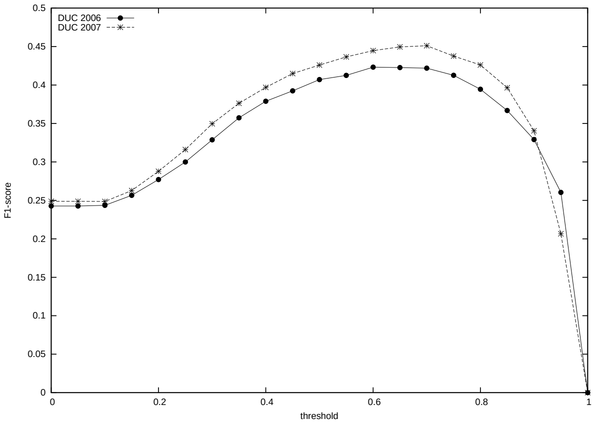

Figure 2: F1-measure for learning which concepts should be included in the summaries

For 2006, the f-measure peaks at 0.42 with a threshold of 0.6. Recall and precision are balanced at this point. In other words, 42% of the concepts within the summaries have been successfully predicted, and 42% of the predicted concepts are summary-worthy. Clearly, there is room for improvement in this stage of the algorithm, but the task is a challenging one. The decision of which concepts to include is subjective, and only 14% (1723 out of 12479) of the training examples are positive. It is encouraging to note the similarities between the 2006 and 2007 results, which indicate that the approach is robust.

Quality of summaries There are several parameters to tune before evaluating the final summarization system, and only the 2006 data was used to do so. The threshold above which concept scores contribute positively to the sentence's score (τ_p) is given by the point illustrated in 2 where F1 score is highest: 0.6. The threshold for scores that penalize the sentence (τ_n) is much lower – 0.2 – because our imperfect classifier often gives low scores to good concepts. The threshold above which two concepts are considered related (τ_r) is set to 0.4, which corresponds to intermediate–or moderately related–on the scale used by Miller and Charles (1991). The threshold for a sentence to be considered related to the topic (τ_t) was raised to 1, meaning they must mention topic concepts directly. The threshold τ_o for considering two sentences as overlapping is 0.55. The weights by which the scores are combined for each sentence were set to: $w_C = w_{KP} = w_{T_{PR}} = 1, w_T = 0.9, w_D = 3$, $w_{Sal} = 0.4$.

DUC (#peers)	ROUGE-2	ROUGE-SU4	BE
2006 (34)	0.08795 (3)	0.14681 (3)	0.04809 (3)
2007 (30)	0.11180 (5)	0.16628 (5)	0.05947 (8)
2008 (71)	0.08522 (17)	0.12451 (16)	0.05133 (22)

Table 1: Summarization results on the development (DUC 2006) and test (DUC 2007 and TAC 2008) data, and the ranking of the performance compared to the peer summaries.

Table 1 shows the system's performance in terms of ROUGE-2, -SU4 and -BE recall on both 2006 and 2007 data. These well-known metrics evaluate the similarity of automatically generated summaries with human produced abstracts through various measures of lexical overlap (Lin, 2004). Our approach ranks 3rd and 5th in DUC 2006 and DUC 2007 respectively in ROUGE-2 scores, compared to the original participants, and 4th and 6th when more recent approaches – (Schilder and Kondadadi, 2008), (Zhang et al., 2008), and (Nastase, 2008) – are considered. The system ranks the same when comparing -SU4 scores.

The system thus tuned was deployed for the update task at TAC 2009. Because the system was developed to produce longer summaries – 250 words vs. 100 words in the update task – we added the sentence compression step to compensate for the bias towards longer sentences. Table 2 shows the system's results with (37) and without (41) sentence compression for the manual and automatic evaluation (in terms of rank, considering the 52 peers, but not the baselines).

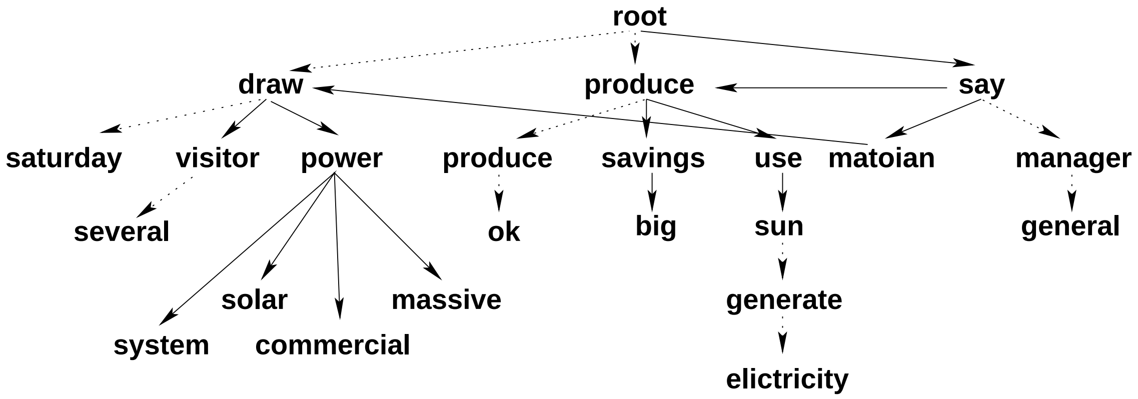

Figure 3: The transformed dependency tree with pruned subtrees

Run ID	Manual	ROUGE-2	ROUGE-SU4	BE
41-A	21	17	15	13
41-B	24	17	19	14
37-A	45	26	25	20
37-B	44	37	29	19

Table 2: Update task results for TAC 2009 – ranks relative to the 52 manually evaluated peers.

Query: Report on the marriage of Camilla Parker Bowles to Prince Charles. Include engagement activities, planning for the wedding, and reaction to the engagement. Do not include Camilla's activities prior to her engagement.

Summary: *Prince Charles' sons William and Harry are "delighted" at news he will marry Camilla Parker Bowles, a spokesman for the heir to the throne's office said Thursday. In a surprise announcement, Britain's heir to the throne, Prince Charles, declared on Thursday that he will marry his lover, Camilla Parker Bowles, in a spring ceremony at Windsor Castle. Charles and Camilla will marry in a civil ceremony, not in the Church of England, and that Parker Bowles will be called Her Royal Highness the Duchess of Cornwall after the marriage, and called the princess consort, not the queen, when Charles becomes king. The wedding is to take place on April 8 in a civil ceremony at Windsor Castle, Charles announced Thursday, but the 57-year-old bride will not become the Princess of Wales – that position having already been more than filled by the prince's late and much-remembered ex-wife, Diana.*

Figure 4: Sample query and summary from TAC 2009

Discussion To analyze the behaviour of the system, let us take a closer look at a good and a bad summary. One of the good summaries in this year's update task is about the marriage of Prince Charles and Camilla Parker Bowles, presented in Figure 4.

The system has identified *civil ceremony*, *ceremony*, *Harry*, *Windsor Castle*, *throne*, *Britain*, *Her Royal Highness*, *princess*, *consort*, *king*, *queen* as pertinent, and this allowed it to select relevant sentences. None of these concepts appear in the query. This clearly illustrates the value of encyclopedic knowledge: such concepts were only deemed related and summary-worthy through consultation with Wikipedia.

We illustrate a bad summary with an example from DUC 2007, about the extent and causes of obesity in the US. It begins promisingly:

Years of emphasizing exercise and healthier diets have done little to stem the growing epidemic of obesity in the United States, with more than 20 percent of adults and up to 15 percent of children now considered obese, a top US health official said Friday.

Unfortunately the remaining sentences are almost exclusively concerned with the health problems that obesity contributes to: e.g. *Obesity is linked to many chronic and life-threatening diseases, including heart disease, diabetes and hypertension. Respiratory disease*, *Diabetes*, and *Cancer* are also discussed. All of these concepts relate strongly to obesity, but are not what the topic asked for. The problem is frustratingly subtle: the system failed to distinguish between the causes of obesity, and what it causes. This illustrates the difficulty of the task, and the importance of the grammatical dependencies and semantic relations between concepts.

Because of its style of scoring sentences based on the concentration of topic-related concepts, the system favours long sentences. The sentence compression module was introduced to counter this bias, but it is a work in progress. Text interpretability is a necessary condition for its informativity, and ungrammatical, incomprehensible sentences fail at delivering any content at all. From this perspective the results are discouraging – most of our compressions had grammatical flaws. A closer inspection of error sources revealed the following:

1. Since the method relies on dependency structures, it is very sensitive to parser errors which were not uncommon: on average top-ranking sentences are considerably longer and thus more difficult to parse. Thus, in many cases dependency structures to be compressed were already corrupted and no grammatical compression would be possible.

2. The hard constraint on the tree size resulted in incomplete sentences. For example, it was not uncommon to get a sentence where a finite verb did not have any arguments because a dependency from the verb would exceed the permitted limit. In the future we would like to replace the general tree size constraint with the one taking into account which dependencies are already present in the tree.

3. Some sentences were grammatical but had a meaning different from that intended in the input because words necessary for correct interpretation were omitted (e.g., omission of a degree adverb like *slightly* in *slightly improved*). Since our method is purely syntax-based, semantic issues have not been addressed yet. However, the constraint framework we adopted makes the system easy to extend with additional constraints and we are planning to do it in the future.

As an illustration to the method in general and to the weakness described in the final point in particular, consider the following sentence from D0903A-A: *Using the sun to generate electricity is producing big energy savings for OK Produce, said general manager Brady Matoian, whose massive commercial solar power system drew several visitors Saturday.*
and its compression: *Using the sun is producing big energy savings said Brady Matoian whose massive commercial solar power system drew visitors.*

Figure 3 presents its transformed dependency tree, which is no longer a tree but a dependency graph, with pruned dependencies marked with dotted lines. Although the compression is grammatical, its meaning appears to be incomplete or even wrong as the quotation has a broader scope (i.e., *big savings for everyone*) than intended by the speaker (i.e., *big savings for OK Produce*).

7 Conclusions

This paper has presented a novel approach for multi-document summarization, which identifies the Wikipedia concepts that are discussed within documents, predicts those that are likely to be found in well-formed summaries, and extracts the relevant sentences accordingly. We have also experimented with a sentence compression module, to reduce the highest scoring sentences to the core that is directly relevant to the user's query.

Further development of the sentence compression module – possibly towards fusing together sentences that share some information but contain novel facts as well – will reduce redundancy and make the summary more concentrated in information relevant to the user.

We have explored the utility of Wikipedia's encyclopedic knowledge for summarization; of being able to identify and relate the people, places, events and ideas mentioned in documents and queries. However there is much room to improve how we take advantage of this resource. There are many features of Wikipedia – for example its category structure, textual content, and the ontologies that have been extracted from it – left to be explored and applied.

Moreover, it makes little sense to use Wikipedia and encyclopedic knowledge exclusively, in isolation from more well known – and more thoroughly investigated – linguistic resources. While Wikipedia can tell us much about specific entities, it has less to say about the prose that ties them together. Simply put, it is not only important what concepts a document, sentence or query talks about, but also what it says about them. This information is largely lost within our current representations, and the generated summaries suffer as a result. We expect that the combination of both linguistic and encyclopedic knowledge will be a very fruitful line of enquiry.

References

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, 15–20 June 2008, pages 807–815.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 22–28 May 2006, pages 449–454.

H.P. Edmundson. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285. Reprinted in *Advances in Automatic Text Summarization*, Mani, I. and Maybury, M.T. (Eds.), Cambridge, Mass.: MIT Press, 1999, pp.21-42.

Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Katja Filippova and Michael Strube. 2008. Dependency tree based sentence compression. In *Proceedings of the 5th International Conference on Natural Language Generation*, Salt Fork, Ohio, 12–14 June 2008, pages 25–32.

Sanda M. Harabagiu. 2004. Incremental topic representations. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 23–27 August 2004, pages 583–589.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 7–12 July 2003, pages 423–430.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for automatic summarization. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany, 31 July – 4 August 2000, pages 495–501.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out Workshop at ACL '04*, Barcelona, Spain, 25–26 July 2004, pages 74–81.

H.P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the ACM 16th Conference on Information and Knowledge Management (CIKM 2007)*, Lisbon, Portugal, 6–9 November, 2007, pages 233–242.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 25–26 July 2004, pages 404–411.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

David Milne and Ian H. Witten. 2008a. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WIKIAI 2008)* Chicago, IL., 13–14 July 2008.

David Milne and Ian H. Witten. 2008b. Learning to link with Wikipedia. In *Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM 2008)*, Napa Valley, CA., 26–30 October, 2008, pages 509–518.

Vivi Nastase. 2008. Topic-driven multi-document summarization with encyclopedic knowledge and activation spreading. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, Hawaii, 25-27 October 2008. To appear.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Mateo, Cal.

G.J. Rath, A. Resnick, and T.R. Savage. 1961. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143. Reprinted in *Advances in Automatic Text Summarization*, Mani, I. and Maybury, M.T. (Eds.), Cambridge, Mass.: MIT Press, 1999, pp.287-292.

Frank Schilder and Ravikumar Kondadadi. 2008. Fastsum: Fast and accurate query-based multi-document summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, 15–20 June 2008, pages 205–208.

Krysta Svore, Lucy Vanderwende, and Christopher Burges. 2008. Enhancing single-document summarization by combining RankNet and third-party sources. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Waikiki, Honolulu, Hawaii, 25-27 October 2008, pages 448–457.

Shiren Ye, Long Qiu, Tat-Seng Chua, and Min-Yen Kan. 2005. NUS at DUC 2005: Understanding documents via concept links. In *Proceedings of the 2005 Document Understanding Conference held at the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, B.C., Canada, 9–10 October 2005.

Jin Zhang, Xueqi Cheng, Gaowei Wu, and Hongbo Xu. 2008. Adasum: an adaptive model for summarization. In *Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM 2008)*, Napa Valley, CA., 26–30 October, 2008, pages 901–910.