

# Topic-Driven Multi-Document Summarization with Encyclopedic Knowledge and Spreading Activation

Vivi Nastase  
EML Research gGmbH  
Heidelberg, Germany  
nastase@eml-research.de

## Abstract

Information of interest to users is often distributed over a set of documents. Users can specify their request for information as a query/topic – a set of one or more sentences or questions. Producing a good summary of the relevant information relies on understanding the query and linking it with the associated set of documents. To “understand” the query we expand it using encyclopedic knowledge in Wikipedia. The expanded query is linked with its associated documents through spreading activation in a graph that represents words and their grammatical connections in these documents. The topic expanded words and activated nodes in the graph are used to produce an extractive summary. The method proposed is tested on the DUC summarization data. The system implemented ranks high compared to the participating systems in the DUC competitions, confirming our hypothesis that encyclopedic knowledge is a useful addition to a summarization system.

## 1 Introduction

Topic-driven summarization reflects a user-based summarization task: from a set of documents derive a summary that contains information on a specific topic of interest to a user. Producing a good summary relies on “understanding” the user’s information request, and the documents to be summarized. It is commonly agreed that the verbal part of a text provides pointers to a much larger body of knowledge we assume the listener has. An American citizen, for example, when told *There will be*

*fireworks on July 4<sup>th</sup>*, understands that there will be a celebration involving fireworks on the occasion of the U.S. Independence Day. Understanding an utterance implies lexical, common-sense and encyclopedic knowledge. Lexical knowledge is usually incorporated in systems through machine readable dictionaries, wordnets or thesauri. Common-sense and encyclopedic knowledge were harder to capture, but recently Wikipedia has opened the possibility of accessing such knowledge on a large scale, and in numerous languages.

To “understand” a user’s information request – one or more sentences or questions (the *topic* of the summary) – summarization systems try to expand it. This will provide later stages of processing with more keywords/keyphrases for retrieving from the documents relevant fragments. In this paper we experiment with Wikipedia for topic expansion. The body of research involving Wikipedia as a source of knowledge is growing fast, as the NLP community finds more and more applications of this useful resource: it is used to acquire knowledge (Suchanek et al., 2007; Auer et al., 2007); to induce taxonomies and compute semantic relatedness (Ponzetto & Strube, 2007b; 2007a); as a source of features for text classification (Gabrilovich & Markovitch, 2006) and for answering questions (Ahn et al., 2004; Katz et al., 2005). The work presented here uses hyperlinks in Wikipedia articles to expand keywords and keyphrases extracted from the query. Ambiguous words are disambiguated using the context provided by the query.

“Understanding” the documents to be summarized implies identifying the entities mentioned, how

they are connected, and how they are related to the entities in the topic. For this, we start again from the topic, and spread an activation signal in a large graph that covers all documents for this topic – nodes are words/named entities in the texts, links are grammatical relations. This way we cross from the topic to the documents, and combine information which is important in the topic with information which is important and relevant in the documents. We take the most highly activated nodes as additional topic expansions, and produce an extractive summary by choosing from the sentences that connect the topic expansion words in the large document graph.

The experiments confirm that Wikipedia is a source of useful knowledge for summarization, and that further expanding the topic within the associated set of documents improves the summarization results even more. We compare the performance of the summarization system to that of participating systems in the DUC competitions. The system we describe ranks 2<sup>nd</sup>, 9<sup>th</sup> and 5<sup>th</sup> in terms of ROUGE-SU4 on the DUC 2005, DUC 2006 and DUC 2007 data respectively.

## 2 Related Work

While the recent exponential increase in the amount of information with which we must cope makes summarization a very desirable tool in the present, summarization is not a novel task. Rath et al. (1961) and Edmundson (1969) have explored extractive summary formation, and have raised important evaluation issues for extractive summaries when compared to several human produced gold standards. Nowadays, summarization methods try to incorporate tools, methodologies and resources developed over the past decades. The NIST organized competitions under the Document Understanding Conferences – DUC (since 2008, Text Analysis Conference (TAC))<sup>1</sup> events provide a forum for the comparison of a variety of approaches, ranging from knowledge poor – Gotti et al. (2007) rely exclusively on a parser, without any additional sources of information – to knowledge rich and complex – GISTexter (Hickl et al., 2007) combines question answering, textual entailment, topic signature modules and a va-

<sup>1</sup><http://duc.nist.gov/>, <http://www.nist.gov/tac>.

riety of knowledge sources for summarization.

The most frequently used knowledge source in NLP in general, and also for summarization, is WordNet (Fellbaum, 1998). Barzilay & Elhadad (1999) use WordNet to model a text’s content relative to a topic based on lexical chains. The sentences intersected by the most and strongest chains are chosen for the extractive summary. Alternative sources for query expansion and document processing have also been explored. Amini & Usunier (2007) use the documents to be summarized themselves to cluster terms, and thus expanding the query “internally”. More advanced methods for query expansion use “topic signatures” – words and grammatically related pairs of words that model the query and even the expected answer from sets of documents marked as relevant or not (Lin & Hovy, 2000; Harabagiu, 2004).

Graph-based methods for text summarization work usually at the level of sentences (Erkan & Radev, 2004; Mihalcea & Tarau, 2004). Edge weights between sentences represent a similarity measure, and a PageRank algorithm is used to determine the sentences that are the most salient from a collection of documents and closest to a given topic. At the word level, Leskovec et al. (2004) build a document graph using subject-verb-object triples, semantic normalization and coreference resolution. They use several methods (node degree, PageRank, Hubs, etc.) to compute statistics for the nodes in the network, and use these as attribute values in a machine learning algorithm, where the attribute that is learned is whether the node should appear in the final summary or not. Annotations for training come from human produced summaries. Mohamed & Rajasekaran (2006) incrementally build a graph for a document collection by combining graph-representations of sentences. Links between entities in a sentence can be *isa* (within an NP) or *related\_to* (between different phrases in a sentence). Nodes and relations are weighted according to their connectivity, and sentence selection for the final summary is based on the most highly connected nodes. Ye & Chua (2006) build an extractive summary based on a concept lattice, which captures in a hierarchical structure co-occurrences of concepts among sentences. Nodes higher in this structure correspond to frequently co-occurring terms, and are

<pre> &lt;topic&gt; &lt;num&gt; D0704A &lt;/num&gt; &lt;title&gt; Amnesty International &lt;/title&gt; &lt;narr&gt; What is the scope of operations of Amnesty International and what are the international reactions to its activities? Give examples of charges lodged by the organization and complaints against it. &lt;/narr&gt; &lt;docs&gt; ... &lt;/docs&gt; &lt;/topic&gt; </pre>	<pre> &lt;topic&gt; &lt;num&gt; D0740I &lt;/num&gt; &lt;title&gt; round-the-world balloon flight &lt;/title&gt; &lt;narr&gt; Report on the planning, attempts and first successful balloon circumnavigation of the earth by Bertrand Piccard and his crew. &lt;/narr&gt; &lt;docs&gt; ... &lt;/docs&gt; &lt;/topic&gt; </pre>
--	---

Figure 1: Sample topics from DUC 2007

assumed to be more representative with respect to the document topic.

Mani & Bloedorn (1999) build a “chronological” graph, in which sentence order is respected and each occurrence of a concept is a separate node. Edges between nodes cover several types of relations: adjacency (ADJ); identity – instance of the same word (SAME); other semantic links, in particular synonymy and hypernymy; PHRASE links connect components of a phrase; NAME indicate named entities; COREF link coreferential name instances. Among other things, they identify regions of the text salient to a user’s query, based on spreading activation starting from query words in this document graph. Spreading activation was introduced in the 60s and 70s to model psychological processes of memory activation in humans (Ross Quillian, 1967; Collins & Loftus, 1975).

In this approach we use Wikipedia as a source of knowledge for related concepts – the texts of hyperlinks in an article describing a concept are taken as its related concepts. The query is further expanded by using spreading activation to move away from the topic in a large graph that covers all documents for a given topic. From the nodes thus reached we select using a PageRank algorithm the ones that are most important in the documents. We study the impact of a decay parameter which controls how far to move from the topic, and the number of highest ranked nodes to be added to the expanded topic. The summary is built based on word associations in the documents’ graph.

### 3 Topic Expansion with Encyclopedic Knowledge or WordNet

In DUC topic-driven multi-document summarization, the topic has a title, an ID that links it to a set of documents, and one or more sentences and/or questions, as illustrated in Figure 1.

Topic processing is done in several steps:

**1. Preprocessing:** Produce the dependency pair representation of the topics using the Stanford Parser<sup>2</sup>. Pairs that have closed-class words are filtered out, and the remaining words are lemmatized<sup>3</sup>. We extract named entities (NEs), as the parser works at the word level. In the dependency pairs we replace an NE’s fragments with the complete NE.

**2a. Query expansion with Wikipedia:** Extract all open-class words and NEs from the topic, and expand them using Wikipedia articles whose titles are these words or phrases.

For each Wikipedia article we extract as related concepts the texts of the hyperlinks in the first paragraph (see Figure 2<sup>4</sup>). The reason for not including links from the entire article body is that apart from the first paragraph, which is more focused, often times hyperlinks are included whenever the under-

<sup>2</sup><http://nlp.stanford.edu/software/lex-parser.shtml>

<sup>3</sup>Using XTAG morphological database <ftp://ftp.cis.upenn.edu/pub/xtag/morph-1.5/morph-1.5.tar.gz>.

<sup>4</sup>The left side shows the first paragraph as it appears on the page, the right side shows the corresponding fragment from the source file, with the annotations specific to Wikipedia.

## Mining

Mining is the extraction of **valuable minerals** or other **geological** materials from the earth, usually (but not always) from an **ore** body, **vein** or (coal) seam. Materials recovered by mining include **bauxite**, **coal**, **copper**, **gold**, **silver**, **diamonds**, **iron**, **precious metals**, **lead**, **limestone**, **magnesite**, **nickel**, **phosphate**, **oil shale**, **rock salt**, **tin**, **uranium** and **molybdenum**. Any material that cannot be grown from **agricultural** processes, or created **artificially** in a **laboratory** or **factory**, is usually mined. Mining in a wider sense comprises extraction of any **non-renewable resource** (e.g. **petroleum**, **natural gas**, or even **water**).

'''Mining''' is the extraction of `[[value (economics)|valuable]]` `[[mineral]]s` or other `[[geology|geological]]` materials from the earth, usually (but not always) from an `[[ore]]` body, `[[vein (geology)|vein]]` or (coal) seam. Materials recovered by mining include `[[bauxite]]`, `[[coal]]`, `[[copper]]`, `[[gold]]`, `[[silver]]`, `[[diamond]]s`, `[[iron]]`, `[[precious metal]]s`, `[[lead]]`, `[[limestone]]`, `[[magnesite]]`, `[[nickel]]`, `[[phosphate]]`, `[[oil shale]]`, `[[Sodium chloride|rock salt]]`, `[[tin]]`, `[[uranium]]` and `[[molybdenum]]`. Any material that cannot be grown from `[[agriculture|agricultural]]` processes, or created `[[Chemical synthesis|artificially]]` in a `[[laboratory]]` or `[[factory]]`, is usually mined. Mining in a wider sense comprises extraction of any `[[non-renewable resource]]` (e.g., `[[petroleum]]`, `[[natural gas]]`, or even `[[fossil water|water]]`).

Extracted related concepts for *mining*:

value (economics), valuable, mineral, geology, geological, ore, vein (geology), vein, coal, bauxite, copper, gold, silver, diamond, iron, precious metal, lead, limestone, magnesite, nickel, phosphate, oil shale, Sodium chloride, rock salt, agriculture, agricultural, Chemical synthesis, artificially, laboratory, factory, non-renewable resource, petroleum, natural gas, fossil water, water.

Figure 2: First paragraph for article *Mining* in the English Wikipedia, and the extracted related concepts.

Word	Wikipedia expansion	WordNet expansion
mining	lead, agricultural, mineral, gold, ore, petroleum, nickel, iron, coal, tin, value, copper, water, bauxite, silver, diamond	production
flight	lift, air	pass, trip, lam, overflight, ballooning, nonstop flight, aviation, soaring, air, flying, solo, break, escape
status	registered	way, situation, mode, position, place, par, need, light, danger, health, state, standing, face, rank, demand, command, control
Southern Poverty Law Center	racism, American, United States, research, civil rights, litigation	–

Table 1: Expanded concepts from DUC 2007 topics, after filtering based on the documents to be summarized.

lying concept appears in Wikipedia, without it being particularly relevant to the current article.

To expand a word (or NE)  $W$  from the query, we search for an article having  $W$  as the title, or part of the title.

1. If one exact match is found (e.g. Southern Poverty Law Center), extract the related concepts for this article.
2. If several exact or partial matches are found, use the larger context of the query to narrow down to the intended meaning. For example, *Turkey* – referring to the country – appears in several topics in the DUC 2007 data. There are multiple entries for “Turkey” in Wikipedia

– for the country, the bird, cities with this name in the U.S. among others. We use a Lesk-like measure, and compute the overlap between the topic query and the set of hyperlinks in the first paragraph (Lesk, 1986). We choose the expansion for the entry with the highest overlap. If the query context does not help in disambiguation, we use the expansions for all partial matches that tie for the highest overlap.

3. If an article with the required name does not exist, the word will not be expanded.

**2b. Query expansion with WordNet:** Extract all nouns and NEs from the topic, and expand them

with hypernyms, hyponyms and antonyms in WordNet 2.0:

1. If an word (or NE)  $W$  from the query corresponds to an unambiguous entry in WordNet, expand that entry.
2. If  $W$  has multiple senses, choose the sense(s) which have the highest overlap with the query. To compute overlap, for a sense we take its expansions (one step hypernyms, hyponyms and antonyms) and the words from the definition.
3. If  $W$  has no senses in WordNet, the word will not be expanded.

**3. Expansion filtering:** Filter the list of related concepts: keep only terms that appear in the document collection for the current topic.

Table 1 includes the expansions obtained from Wikipedia and from WordNet respectively for a number of words in topics from the DUC 2007 collection. *mining* is a specific activity, involving a limited set of materials. While such connections cannot be retrieved through hypernym, meronym or other semantic relations in WordNet, they are part of encyclopedic knowledge, and can be found in Wikipedia. *flight* is a more general concept – there are specific types of flight, which appear as hyponyms in WordNet, while in Wikipedia it is more generally described as the motion of an object through air, which does not provide us with interesting related concepts. *status* is a very general concept, and rather vague, for which neither WordNet nor Wikipedia can provide very useful information. Finally, Wikipedia is rich in named entities, which are not in the scope of a semantic lexicon. WordNet does contain named entities, but not on the scale on which Wikipedia does.

For the 45 topics from DUC 2007, the expansion with Wikipedia generated 1054 additional words, while with WordNet 2510. This difference comes from the fact that with Wikipedia it is mostly the NEs that are expanded, whereas with WordNet the common nouns, which are more numerous in the topics. The overlap between the two sets of expansions is 48 words (0.046 relative to Wikipedia expansions, 0.019 relative to WordNet).

## 4 Topic Expansion with Spreading Activation and PageRank

Concepts related to the ones in the topic provide a good handle on the documents to summarize – they indicate parts of the document that should be included in the summary. It is however obvious that the summary should contain more than that, and this information comes from the documents to be summarized. Amini & Usunier (2007) have shown that expanding the query within the set of documents leads to good results. Following this idea, to find more relevant concepts we look for words/NEs which are related to the topic, and at the same time important in the collection of documents for the given topic. The methods described in this section are applied on a large graph that covers the entire document collection for one topic. The documents are processed in a similar way to the query – parsed with the Stanford Parser, output in dependency relation format, lemmatized using XTag’s morphological data file. The graph consists of nodes corresponding to lemmatized words and NEs in the documents, and edges corresponding to grammatical dependency relations.

### 4.1 Spreading Activation

To find words/NEs related to the topic we spread an activation signal starting from the topic words and their expansions (in a manner similar to (Mani & Bloedorn, 1999), and using an algorithm inspired by (Anderson, 1983)), which are given a node weight of 1. As we traverse the graph starting from these nodes, the signal is propagated by assigning a weight to each edge and each node traversed based on the signal strength. The signal strength diminishes with the distance from the node of origin depending on a signal decay parameter, according to the formula:

$$\begin{aligned}
 w_n(N_0) &= 1; \\
 s_t &= (1 - decay) * \frac{w_n(N_t)}{Out(N_t)}; \\
 w_n(N_{t+1}) &= s_t; \\
 w_e(N_t, N_{t+1})_{t+1} &= w_e(N_t, N_{t+1})_t + s_t;
 \end{aligned}$$

where  $N_t$  is the current node;  $N_{t+1}$  is the node we are moving towards;  $w_n(N_t)$  is the weight of node  $N_t$ ;  $s_t$  is the signal strength at step  $t$ ;  $Out(N_t)$

Topic	Topic expanded words	Top ranked nodes
D0738 What is the status of mining in central and South America? Include obstacles encountered.	<i>status, registered, South America, central, 1998, obstacle, mining, lead, agricultural, mineral, gold, ore, petroleum, nickel, iron, coal, tin, value, copper, water, bauxite, silver, diamond, include, encounter</i>	company, dollar, project, sector, iron, mine, silver, percent, big, value, industry, source, overturn, regulate, link, official, decree, financing, expert, firm, activity, estimate, state, For Peru, Peru, third, already, top, 12th, creation, ton
D0717 Describe the various lawsuits against American Home Products which resulted from the use of fenfluramine, also known as Pondimin, and half of the diet drug combination called "fen-phen".	<i>combination, set, half, American Home Products, know, fenfluramine, phentermine, obesity, release, dexfenfluramine, use, United States, Wal-Mart, fen, describe, diet, call, drug, drugs, medication, patients, medicine, lawsuit, right, court, damages, defendant, plaintiff, also, various, Pondimin, result</i>	drug, market, company, settle, reduce, claim, American Home Products, make, cause, seek, cover, people, allow, agree, dismiss, other, sue, case, Pondimin, state, link, million, award, user, estimate, thousand, file, think, note, damages, Harris County

Table 2: Top ranked nodes after expanding the topic with spreading activation and PageRank

is the number of outgoing edges from node  $N_t$ ;  $w_e(N_t, N_{t+1})_t$  is the weight of the edge between  $N_t$  and  $N_{t+1}$  at time  $t$  (i.e., before actually traversing the edge and spreading the activation from  $N_t$ );  $w_e(N_t, N_{t+1})_{t+1}$  is the weight of the edge after spreading activation. The weight of the edges is cumulative, to gather strength from all signals that pass through the edge. Activation is spread sequentially from each node in the (expanded) topic.

The *decay* parameter is used to control how far the influence of the starting nodes should reach – the lower the decay, the farther the signal can reach.

## 4.2 PageRank

The previous step has assigned weights to edges in the graph, such that higher weights are closer to topic and/or topic expanded words. After this initialization of the graph, we run a PageRank algorithm (Brin & Page, 1998) to determine more important nodes. By running this algorithm after initializing the graph edge weights, from the nodes that are closer to topic and topic expanded words we boost those that are more important in the documents.

The starting point of the PageRank algorithm is the graph with weighted edges obtained in the previous step. The node weights are initialized with 1 (the starting value does not matter). Analysis of the documents graph for several topics has revealed that there is a large highly interconnected structure, and many disconnected small (2-3 nodes) fragments.

Page Rank will run on this dense core structure. The PageRank algorithm is guaranteed to converge if the graph is aperiodic and irreducible (Grimmett & Stirzaker, 1989). Aperiodicity implies that the greatest common divisor of the graph’s cycles is 1 – this condition is met. Irreducibility of the graph means that it has no leaves, and there are no two nodes with the same set of neighbours. The remedy in such cases is to connect each leaf to all other nodes in the graph, and conflate nodes with the same set of neighbours.

Once the graph topology meets the PageRank convergence conditions, we run the algorithm. The original formula for computing the rank of a node at each iteration step is:

$$PR(n_i) = \frac{1-d}{N} + d \sum_{n_j \in Adj_{n_i}} \frac{PR(n_j)}{Out(n_j)}$$

where  $n_i$  is a node,  $d$  is the damping factor (usually  $d = 0.85$  and this is the value we use as well),  $N$  is the number of nodes in the graph,  $PR(n_i)$  is the rank of node  $n_i$ ,  $Adj_{n_i}$  is the set of nodes adjacent to  $n_i$ , and  $Out(n_j)$  is the number of outgoing edges from  $n_j$  (our graph is non-directed, so this number is the total number of edges with one end in  $n_j$ ). We adjust this formula to reflect the weights of the edges, and the version used is the following:

$$PR(n_i) = \frac{1-d}{N} + d \sum_{n_j \in Adj_{n_i}} PR(n_j)w_{out}(n_j);$$

Expansion	ROUGE-2	ROUGE-SU4	BE
none	0.09270 (0.08785 - 0.09762)	0.14587 (0.14019 - 0.1514)	0.04958 (0.04559 - 0.05413)
WN (with WSD)	0.09494 (0.09086 - 0.09900)	0.15295 (0.14897 - 0.15681)	0.04985 (0.04606 - 0.05350)
WN (no WSD)	0.09596 (0.09189 - 0.09990)	0.15357 (0.14947 - 0.15741)	0.05173 (0.04794 - 0.05550)
Wiki	<b>0.10173</b> (0.09721 - 0.10608)	<b>0.15725</b> (0.15345 - 0.16130)	<b>0.05542</b> (0.05125 - 0.05967)
WN (no WSD) + Wiki	0.09604 (0.09228 - 0.09980)	0.15315 (0.14923 - 0.15694)	0.05292 (0.04912 - 0.05647)

Table 3: Comparison of topic expansion methods with 95% confidence intervals.

$$w_{out}(n_j) = \sum_{n_k \in Adj_{n_j}} w_e(n_k, n_j)$$

In Table 2 we show examples of top ranked nodes for several topics, extracted with this algorithm. The words in italics are keywords/phrases from the topic query, and the top ranked nodes are listed in decreasing order of their rank.

## 5 Summarization

The summarization method implemented is based on the idea that the entities or events mentioned in the query are somehow connected to each other, and the documents to be summarized contain information that allows us to make these connections. We use again the graph for all the documents in the collection related to one topic, built using the dependency relation representation of the texts. The nodes in this graph are words/NEs, and the links are grammatical relations.

We extract from this graph the subgraph that covers connections between all open class words/NEs in the topic or expanded topic query. Each edge in the extracted subgraph corresponds to a grammatical relation in a sentence of a document. We collect all sentences thus represented in the subgraph, and rerank them based on the number of edges they cover, and the occurrence of topic or expanded topic terms. We use the following formula to compute a sentence score:

$$\begin{aligned} Score(S) = & \textit{topicWords} * w_{word} \\ & + \textit{expandedWords} * w_{expandedWord} \\ & + \textit{topRankedWords} * w_{topRankedWord} \\ & + \textit{edgesCovered} * w_{subgraphEdge} \\ & + \textit{depRelation} * w_{depRelation} \end{aligned}$$

$w_{word}$ ,  $w_{expandedWord}$ ,  $w_{topRankedWord}$ ,  $w_{subgraphEdge}$  and  $w_{depRelation}$  are weight parameters that give different importance to exact

words from the topic, expanded words, top ranked words and edges covered in the extracted subgraph. During all experiments these parameters are fixed.<sup>5</sup>

To form the summary we traverse the ranked list of sentences starting with the highest ranked one, and add sentences to a summary, or delete from the existing summary, based on a simple lexical overlap measure. We stop when the desired summary length is reached – for DUC 2005–2007, 250 words (last sentence may be truncated to fill the summary up to the allowed word limit).

## 6 Evaluation

Experiments are run on DUC 2007 main summarization task data, for the last experiment we used the DUC 2005 and DUC 2006 data as well. Performance is evaluated in terms of ROUGE-2, ROUGE-SU4 and BE recall, following the methodology and using the same parameters as in the DUC summarization events.

We analyze several types of topic expansion: no expansion, WordNet, Wikipedia, and within document collection expansion using spreading activation and Page Rank. The spreading activation method has several parameters whose values must be determined.

We first compare the summaries produced with no topic expansion, WordNet (WN) and Wikipedia (Wiki) respectively. Table 3 shows the results in terms of ROUGE and BE recall on the DUC 2007 (main) data. Word sense disambiguation (WSD) for expansion with WordNet did not work very well,

<sup>5</sup>The values used were set following a small number of experiments on DUC 2007 data, as the purpose was not to tune the system for best performance, but rather to study the impact of more interesting parameters, in particular expansion type, decay and node ranking. The values used are the following:  $w_{word} = 5$ ,  $w_{expandedWord} = 2.5$ ,  $w_{topRankedWord} = 0.5$ ,  $w_{subgraphEdge} = 2$ ,  $w_{depRelation} = 0$ .

as evidenced by the lower results for disambiguated expansion (WN with WSD) compared to the non-disambiguated one. A better disambiguation algorithm may reverse the situation. Expanding a topic only with Wikipedia hyperlinks gives the best results. At the document level, the results are not as clear cut. Figure 3 shows a comparison in terms of ROUGE-SU4 recall scores at the document level of the Wikipedia and WN (no WSD) expansion methods, sorted in increasing order of the Wikipedia-based expansion scores. The points are connected to allow the reader to follow the results for each method.

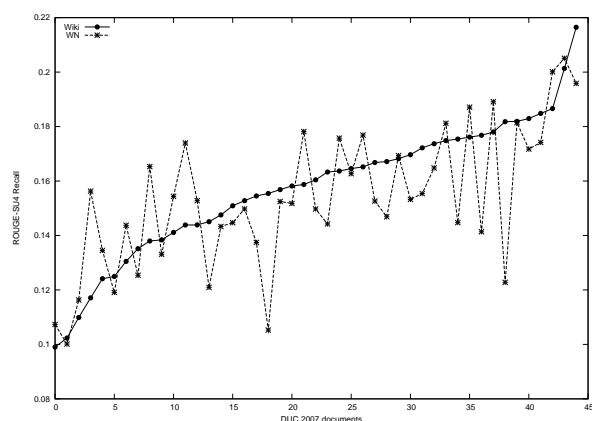


Figure 3: Comparison of Wikipedia and WN ROUGE-SU4 per-document recall results.

Because the overlap between Wikipedia and WordNet expanded queries was very low, we expected the two types of expansion to be complementary, and the combination to give better results than either expansion by itself. An analysis of results for each document with the three expansion methods – Wikipedia, WordNet, and their combination – showed that the simple combination of the expanded words cannot take advantage of the situations when one of the two methods performs better. In future work we will explore how to detect, based on the words in the query, which type of expansion is best, and how to combine them using a weighting scheme.

We choose the best configuration from above (Wikipedia expansion), and further expand the query through spreading activation and PageRank. This new type of expansion has two main parameters which influence the summarization outcome: number of top ranked nodes to add to the topic expansion,

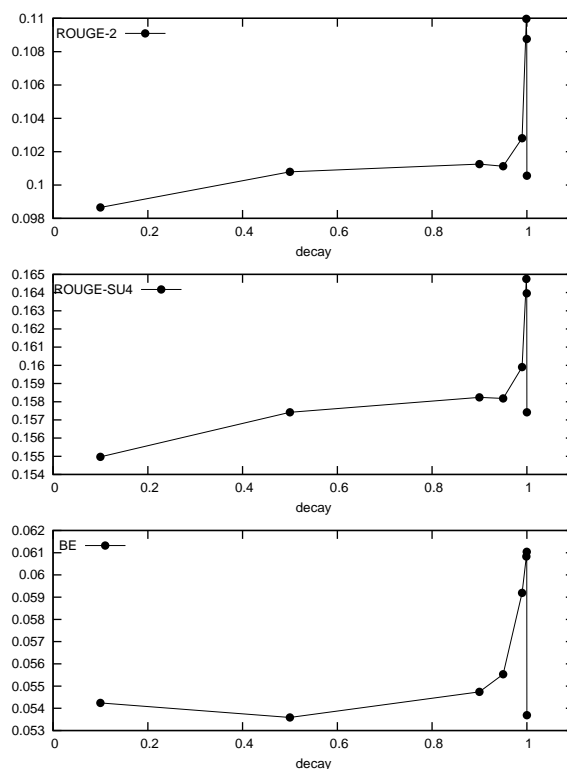


Figure 4: Impact of signal decay in spreading activation on summarization performance.

and the decay of the spreading activation algorithm.

The decay parameter determines how far the influence of the starting nodes (words from query or Wikipedia-expanded query) should be felt. The results in Figure 4 – for decay values 0.1, 0.5, 0.95, 0.99, 0.999, 0.9999, 1 – indicate that faster decay (reflected through a higher decay value) keeps the summary more focused around the given topic, and leads to better results.<sup>6</sup> For a high enough decay – and eventually a decay of 1 – the weights of the edges become extremely small, and due to real number representation in memory, practically 0. In this situation PageRank has no effect, and all nodes have the same rank.

We fix the decay parameter to 0.9999, and we study the impact of the number of top nodes chosen after ranking with PageRank. Figure 5 shows the results when the number of top ranked nodes chosen

<sup>6</sup>During this set of experiments all other parameters are fixed, the number of top ranked nodes added to the topic expansion is 30.

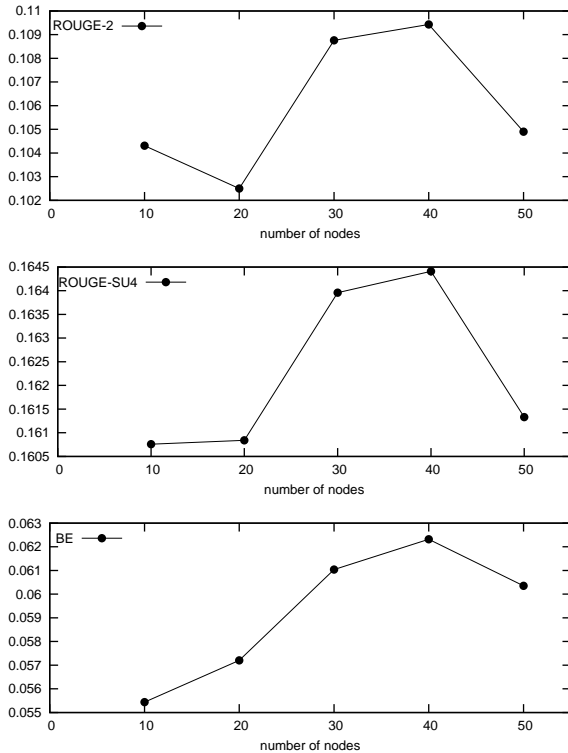


Figure 5: Impact of the number of top ranked nodes added to the expanded topic on summarization performance.

varies. Adding highly ranked nodes benefits the performance of the system only up to a certain limit. From the values we tested, the best results were obtained when adding 40 nodes to the expanded topic.

The best system configuration from the ones explored<sup>7</sup> is run on the DUC 2005, 2006 and 2007 (main) data. The performance and rank (in parentheses) compared to participating systems is presented in Table 4.

DUC	ROUGE-2	ROUGE-SU4	BE
2005 (32)	0.07074 (3)	0.13002 (2)	–
2006 (35)	0.08091 (11)	0.14022 (9)	0.04223 (11)
2007 (32)	0.11048 (6)	0.16479 (5)	0.06250 (5)

Table 4: System performance (and rank) on the DUC 2005, 2006 and 2007 (main) data. The number in parenthesis after the DUC year indicates the number of competing systems.

<sup>7</sup>Wikipedia expansion + 40 top nodes after spreading activation and PageRank, decay = 0.9999,  $w_{expandedWord} = 3.5$ ,  $w_{depRelation} = 1$ , the other parameters have the same values as before.

## 7 Conclusions

The experiments conducted within the summarization framework of the Document Understanding Conference have confirmed that encyclopedic knowledge extracted from Wikipedia can benefit the summarization task. Wikipedia articles are a source of relevant related concepts, that are useful for expanding a summarization query. Furthermore, including information from the documents to be summarized by choosing relevant concepts – based on closeness to topic keywords and relative importance – improves even more the quality of the summaries, judged through ROUGE-2, ROUGE-SU4 and BE recall scores, as it is commonly done in the DUC competitions. The topic expansion methods explored lead to high summarization performance – ranked 2<sup>nd</sup>, 9<sup>th</sup> and 5<sup>th</sup> on DUC 2005, 2006 and 2007 respectively according to ROUGE-SU4 scores – compared to (more than 30) DUC participating systems.

The graph representation of the documents is central to the summarization method we described. Because of this, we plan to improve this representation by collapsing together coreferential nodes and clustering together related concepts, and verify whether such changes impact the summarization results, as we expect they would.

Being able to move away from the topic within the set of documents and discover new relevant nodes is an important issue, especially from the point of view of a new summarization style – updates. In update summaries the starting point is a topic, which a summarization system must track in consecutive sets of documents. We can adjust the spreading activation parameters to how far a new set of documents is from the topic. Future work includes testing the spreading activation and page ranking method in the context of the update summarization task and exploring methods of extracting related concepts from the full text of Wikipedia articles.

**Acknowledgments** This work was funded by the Klaus Tschira Foundation, Heidelberg, Germany. We thank the anonymous reviewers for insightful comments and suggestions.

## References

- Ahn, D., V. Jijkoun, G. Mishne, K. Müller, M. de Rijke & S. Schlobach (2004). Using Wikipedia at the TREC QA track. In *Proc. of TREC-13*.
- Amini, M. R. & N. Usunier (2007). A contextual query expansion approach by term clustering for robust text summarization. In *Proc. of DUC-07*.
- Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behaviour*, 22:261–295.
- Auer, S., C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak & Z. Ives (2007). DBpedia: A nucleus for a Web of open data. In *Proc. of ISWC 2007 + ASWC 2007*, pp. 722–735.
- Barzilay, R. & M. Elhadad (1999). Using lexical chains for text summarization. In I. Mani & M. T. Maybury (Eds.), *Advances in Automatic Text Summarization*, pp. 111–121. Cambridge, Mass.: MIT Press.
- Brin, S. & L. Page (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.
- Collins, A. M. & E. F. Loftus (1975). A spreading-activation theory of semantic processing. *Psychological Review*, (82):407–428.
- Edmundson, H. (1969). New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.
- Erkan, G. & D. R. Radev (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.
- Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.
- Gabrilovich, E. & S. Markovitch (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proc. of AAAI-06*, pp. 1301–1306.
- Gotti, F., G. Lapalme, L. Nerima & E. Wehrli (2007). GOFASum: a symbolic summarizer for DUC. In *Proc. of DUC-07*.
- Grimmett, G. & D. Stirzaker (1989). *Probability and Random Processes*. Oxford University Press.
- Harabagiu, S. (2004). Incremental topic representations. In *Proc. of COLING-04*, pp. 583–589.
- Hickl, A., K. Roberts & F. L. C. C. Lacatusu (2007). LCC's GISTexter at DUC 2007: Machine reading for update summarization. In *Proc. of DUC-07*.
- Katz, B., G. Marton, G. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller, O. Uzuner & A. Wilcox (2005). External knowledge sources for Question Answering. In *Proc. of TREC-14*.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. of ACS-D-86*, pp. 24–26.
- Leskovec, J., M. Grobelnik & N. Milic-Frayling (2004). Learning sub-structures of document semantic graphs for document summarization. In *Proc. of LinkKDD-04*.
- Lin, C.-Y. & E. Hovy (2000). The automated acquisition of topic signatures for automatic summarization. In *Proc. of COLING-00*, pp. 495–501.
- Mani, I. & E. Bloedorn (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35–67.
- Mihalcea, R. & P. Tarau (2004). TextRank: Bringing order into texts. In *Proc. of EMNLP-04*, pp. 404–411.
- Mohamed, A. A. & S. Rajasekaran (2006). Query-based summarization based on document graphs. In *Proc. of DUC-06*.
- Ponzetto, S. P. & M. Strube (2007a). Deriving a large scale taxonomy from Wikipedia. In *Proc. of AAAI-07*, pp. 1440–1445.
- Ponzetto, S. P. & M. Strube (2007b). Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.
- Rath, G., A. Resnick & T. Savage (1961). The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143.
- Ross Quillian, M. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioural Science*, 12(5):410–430.
- Suchanek, F. M., G. Kasneci & G. Weikum (2007). YAGO: A core of semantic knowledge. In *Proc. of WWW-07*, pp. 697–706.
- Ye, S. & T.-S. Chua (2006). NUS at DUC 2006: Document concept lattice for summarization. In *Proc. of DUC-06*.