Estimating Importance Features for Fact Mining

(With a Case Study in Biography Mining)

Sisay Fissaha Adafre and Maarten de Rijke
ISLA, University of Amsterdam
sfissaha,mdr@science.uva.nl

Abstract

We present a transparent model for ranking sentences that incorporates topic relevance as well as an aboutness and importance feature. We describe and compare five methods for estimating the importance feature. The two key features that we use are graph-based ranking and ranking based on reference corpora of sentences known to be important. Independently those features do not improve over the baseline, but combined they do. While our experimental evaluation focuses on informational queries about people, our importance estimation methods are completely general and can be applied to any topic.

1 Introduction

Over 60% of web search queries are informational in nature, and so-called undirected queries form a significant subclass of these [18]. Users issuing an undirected informational query about X want to learn something/everything about X; as [18] put it, "such queries might be interpreted as "Tell me about X"." We propose to answer such queries by returning a short list of important facts. More specifically, given a document collection, we collect sentences about X, which we then classify as important or not important.

Our main aim is to propose and compare algorithms for identifying vital sentences from a text corpus. We propose a simple and transparent schema for sentence ranking algorithms based on (1) ranking the passage (containing the sentence) with respect to its relevancy for a given topic, (2) the extent to which the sentence discusses the topic, and (3) an importance feature for the sentence. Our main research question concerns this importance feature: how can we estimate it? We explore the impact of two main components. First, the use of graph-based sentence ranking as a way of capturing the intuition that importance is a *global* notion. Second, the use of a reference corpus consisting of sentences (known or assumed to be important) about entities of the same category as the entity for which we are seeking important sentences. Our main contribution is an importance estimation method—plus an evaluation of this method—that combines a corpus-based approach to capturing the knowledge encoded in sentences known to be important with a graph-based method for ranking sentences.

To make matters concrete in our experimental evaluation, we restrict the evaluation to finding important facts about a person—our methods, however, are completely general, and can be applied to any topic. In the people case, we take a sentence to be important if it specifies "identifying" properties of the person and/or important events in his or her life time. These include biographical details (date of birth, death, etc.); education; occupation; start or end of an important event; important achievements; marriage related facts.

The remainder of the paper is organized as follows. Section 2 defines the task in detail. In Section 3 we introduce a schema for importance ranking, and provide descriptions of methods for estimating importance features. Section 4 presents our experimental set-up (including the use of Wikipedia as a reference corpus) and the results of experiments aimed at comparing the effectiveness of various importance features. We discuss related work in Section 5 and conclude in Section 6.

2 The Task

The task we address in this paper is to identify sentences that are most important for the topic under consideration. In other words, our task is to identify, in a general newspaper text corpus, definition-like sentences for the topic at hand. In the experimental evaluation in this paper (Section 4), we restrict ourselves to finding important information about people so as to remain focused. In that case, importance of a sentence is judged from the perspective of biographical importance. Consider an example, where the person of interest is William H. McNeill. The sentences below are marked "V" (vital) or "O" (for okay, meaning not vital but still relevant); no sentence is marked "non-relevant."

V “William H. McNeill” (born 1917, Vancouver, British Columbia) is a Canadian historian.

V He is currently Professor Emeritus of History at the University of Chicago.

V McNeill’s most popular work is “The Rise of the West”.

O The book explored human history in terms of the effect of different old world civilizations on one another, and especially the dramatic effect of western civilization on others in the past 500 years.

O It had a major impact on historical theory, especially in distinction to Oswald Spengler’s view of distinct, independent civilizations.

V McNeill is the son of theologian James T. McNeill and the father of historian, J.R. McNeill.

O His most recent book, with his son, is “The Human Web”.

The vital sentences describe the person directly (informing us about “identifying” properties), while the others describe a book he wrote. In reply to the question “Tell me about William H. McNeill,” the first type is clearly preferred.

Solutions to the task we describe can be used to improve information access by providing a focused summary in response to undirected informational queries about a topic. Second, it can be used as a first step in generating structured information about a given topic, where subsequent processing would organize the material in a coherent manner. Our task is similar to the task assessed at the WiQA pilot at CLEF 2006 [14]: there, the idea is to collect important (new) snippets (from Wikipedia pages) on a topic t that supplement an existing Wikipedia page on t. We, in contrast, are interested in harvesting important snippets from arbitrary sources (news paper corpora, the web, $\ldots$) and do not want to assume that we have a Wikipedia page as a starting point nor do we want to address novelty checking/duplicate detection in this paper. Finally, our task also differs from biography-related tasks considered at DUC 2004 or the TREC QA track. Ours is a sentence retrieval task, where the sentence is the unit of retrieval, and not a smaller unit, and we are not limited to biographies. As a consequence, we need to generate a new test set which allows for a proper evaluation of the methods to be introduced below; see Section 4.

3 Estimating Importance

In this section we describe our algorithms for identifying important sentences. We provide a high level overview, and follow with descriptions of methods for estimating importance features.

3.1 Overview

We propose a group of methods for ranking sentences by importance, based on the following formula. Given a topic t (of category c),[1] we estimate the importance of sentence s for the topic t

$$\mathrm{score}(s, t, c) \propto \mu(s) \cdot \lambda(s) \cdot p(t|passage_s). \tag{1}$$

Here, $passage_s$ is the (unique) passage containing s and $p(t|passage_s)$ is the topic likelihood. The features $\lambda(s)$ and $\mu(s)$ allow us to incorporate two key features:

- the extent to which the sentence s discusses the topic t ($\lambda(s)$), and
- the importance of s for the topic t ($\mu(s)$).

To remain focused, we zoom in on different ways of expressing the importance feature $\mu(s)$, taking reasonably performing baseline settings for passage retrieval and the $\lambda(s)$ parameter; see Section 3.2. Figure 1 summarizes the methods we consider for importance estimation. Some (viz. M1, M2, M4, M5) use a *reference corpus* consisting of *reference sentences*. This corpus is a collection of documents, that depends on the category of the topic. Our assumption is that there is a high degree of content similarity among sentences describing entities in the same category. We operationalize this idea as follows. For a given category c, we create a reference corpus C_c and use this to rank sentences that are about an object in the category c and that have been harvested from, say, a newspaper corpus as potentially important sentences. We refer to sentences in the reference corpus as *reference* sentences, and to the harvested sentences that we aim to rank with respect to importance for a given topic as *target* sentences. The creation of the reference corpora used in this paper is described in Section 4.3.

[1] Determining the category of a person can be set up as a classification task, see e.g., [11]; while interesting, this is outside the focus of our research here.

Not graph-based (Subsection 3.3)

M0 The baseline: rank target sentences according to the retrieval status value of the passages they are contained in.

M1 Word overlap between target sentences and reference corpus.

M2 Likelihood of a target sentence given the reference corpus.

Graph-based (Subsection 3.4)

M3 Use PageRank on a graph of sentences induced by inter-sentence similarity.

M4 Use PageRank on a sentence graph induced by similarity to sentences in a reference corpus.

M5 Use PageRank on a sentence graph induced by similarity to sentences in a reference corpus with explicit representation of important lexical items.

Figure 1: Six methods for estimating sentence importance.

3.2 Passage Retrieval and Sentence Extraction

For passage retrieval we used fixed length non-overlapping passages of 400 characters. The topic likelihood $p(t|passage_s)$ in Eq. 1 is determined by taking the product over all the terms in the topic, such that

$$p(t|passage_s) = \prod_{t' \in t} p(t'|passage_s)^{n(t',t)}$$

(with $n(t', t)$ being the number of times t' occurs in t); $p(t'|passage_s)$ is the maximum likelihood estimate of t'. For efficiency purposes, we consider at most 200 passages per topic for further processing. The passages retrieved are split into sentences; sentences that actually contain a mention of the topic are given a strong—almost Boolean—boost. Specifically, we take the feature $\lambda(s)$ in (1) to be $1/p(t|passage_s)$ if t occurs in s, and 1 otherwise.

3.3 Non-graph Based Importance Estimation

We now discuss three methods of estimating the importance feature $\mu(s)$ in (1). The methods we discuss here avoid the usage of graphs, but, except the baseline estimator, they do make use of a reference corpus: the basic assumption is that there is a high degree of content similarity among sentences describing important facts for a given category of entities; these lexical features can be captured by creating a corpus of facts known (or assumed) to be important for the category.

M0: Baseline The baseline approach for estimating sentence importance simply sets $\mu(s)$ in (1) to 1. Hence, the overall score of a sentence in (1) is determined by the retrieval status value of the passage to which it belongs. Though simplistic, this importance estimation method has been shown to perform well on the (different but) related task of answering definition questions at TREC 2003 [24].

M1: Word Overlap Here, and in method M2 below, a sentence is deemed to be important if it "resembles" a reference corpus of facts assumed to be important. In M1 "resemblance" is computed using word overlap. Specifically, given a topic t (with category c) and a sentence s, the importance feature $\mu(s)$ is the fraction of words in s that occur in C_c, the reference corpus for the category c. (The creation of the reference corpora is detailed in Subsection 4.3.)

We use the Jaccard coefficient (JC) as our word overlap similarity metrics; we compute JC between each possible pair of target sentence and reference sentence, and per target sentence, we retain the maximum score attained. Target sentences whose scores fall below an empirically determined threshold are discarded.

M2: Unigram Language Modeling As in method M1, the feature $\mu(s)$ is determined by "resemblance" to a reference corpus. Here, however, we use a language modeling approach. Given a topic t (with category c), we take $\mu(s)$ to be the likelihood of the target sentence given the reference corpus. That is, $\mu(s) = p(s|C_c)$, where the latter is the product of the smoothed likelihood estimates of the terms occurring in s: $\prod_{t' \in s} p(t'|C_c)^{n(t',s)}$, where

$$p(t'|C_c) = \lambda p_{ml}(t'|C_c) + (1 - \lambda) p(t'|W),$$

and $p_{ml}(t'|C_c)$ is the maximum likelihood estimate based on C_c and $p(t'|W)$ is the background probability estimate based on the Wikipedia corpus.

3.4 Graph-based Importance Estimation

We now describe three graph-based methods for estimating importance features. The use of graph-based methods is motivated by the intuition that importance is a *global* notion, that relates all target sentences, and this is exactly what graph-based methods allow us to achieve. By creating a graph of sentences (or of lexical items and sentences), target sentences can receive "support" from other sentences (or lexical items) and they can "pass on" some of this support in a recursive manner. In this way, some sentences may be able to amass a lot of support—these are the important ones. Let us make this precise.

The three graph-based methods differ in the graphs that they use, but they all use the same weighted PageRank algorithm.

M3: Generic Graph-Based Estimation Here we describe a method that has been used for ranking different types of documents, most importantly web documents. We implemented a simple graph-based ranking method originally proposed by [10] and [17]. We create a graph by connecting target sentences that have a similarity score (for which we use the Jaccard coefficient) above a certain threshold, resulting in a configuration such as Figure 2(Left). Using this graph-configuration, we set $\mu(s)$ to be a weighted PageRank of s, as computed in (2):

$$\mu(s) = PR(s) = \frac{d}{N} + (1-d) \cdot \sum_{u \in adj(s)} \frac{JC\left(s,u\right)}{\sum_{v \in adj(u)} JC\left(v,u\right)} \cdot PR\left(u\right), \qquad (2)$$

where d is the PageRank 'damping factor' which is set to 0.85, N is the total number of nodes in the graph, u and v represent nodes in the graph, and $JC\left(u,v\right)$ is the Jaccard coefficient between u and v [7].

M4: Weighted Graph-Based Estimation We now consider a more advanced graph-based importance estimation method by bringing in "weighted support" from the reference corpus. Specifically, we construct a graph structure of the type shown in Figure 2(Center): each reference sentence is the source of a weighted directed link to target sentences, which is defined as follows. Target sentences with a similarity score greater than zero receive a link; the weight attached to such a link is the Jaccard coefficient between the two sentences. We can think of this as follows: reference sentences "vote" for target sentences, with voting weights depending on their similarity scores. Note that in M4 we link reference sentences to target sentences, but we do not link target sentences to each other—unlike M3, where sentences about a topic interlinked based on their similarity scores. When available, M4 uses an additional graph feature. The sentences in the reference corpus may come equipped with a graph structure, derived, say, from the layout

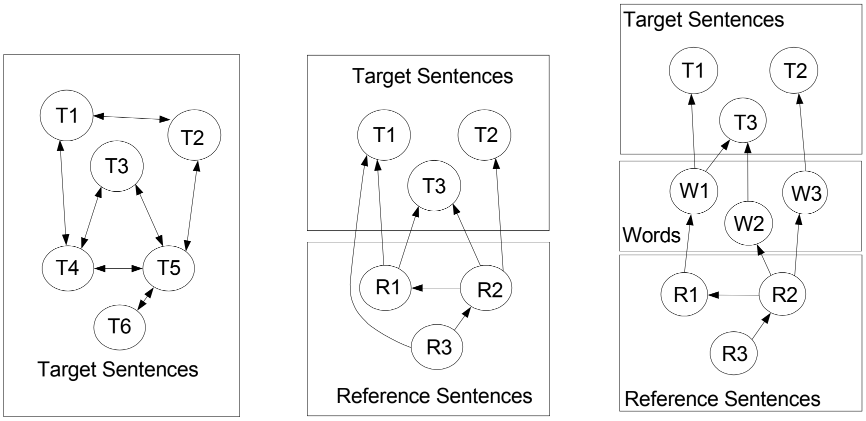

Figure 2: (Left) Generic importance estimation according to M3. (Center) Graph-based importance estimation according to M4. (Right) Graph-based importance estimation according to M5.

or link structure of the documents from which they originate. If this additional information is present, M4 incorporates it without problem.

Once we have the graph configuration described above, the computation of the importance feature $\mu(s)$ of target sentences s is based on the weighted PageRank algorithm as with method M3.

M5: Weighted Graph-Based Estimation with Explicit Lexical Items While M4 incorporates information from a reference corpus, it does so at the level of sentences, which may contain extraneous information. Estimation method M5 enriches the graph used by M4 by explicitly representing lexical items that are important in the reference corpus. E.g., in the biography domain, likely indicators of importance include general terms such as “born,” “died,” “famous,” etc., as well as occupation-specific ones such as, e.g., “architect,” “designed,” “buildings,” for architects.

We incorporate important lexical items by introducing a minor modification of the structure proposed in Figure 2(Center). Reference sentences “vote” for the important lexical items they contain, which in turn “vote” for the target sentences in which they are contained; see Figure 2(Right). The voting is done by passing scores to the neighboring nodes in the graph, which is done by the weighted PageRank algorithm (2). The PageRank algorithm computes the score of a node in the graph based on the scores of its neighbors. That means, a lexical item in the middle layer receives higher scores if it appears in a larger number of reference sentences, and a target sentence in the upper layer receives higher scores if it contains more lexical items in the middle layer.

Once we have obtained the graph configuration described above, the computation of the importance feature $\mu(s)$ of target sentences s is based on the PageRank algorithm as with methods M3 and M4.

4 Experimental Evaluation

To compare and analyze our importance estimation methods we conduct a number of experiments. Below, we detail our experimental set-up, including the creation of reference corpora, and then present our results and error analysis.

4.1 Research Questions

We report on experiments aimed at answering the following questions: Does the use of reference corpora help in improving importance estimation? Does graph-based estimation methods outperform non-graph-based methods? And, does the additional representation of important lexical items help improve importance estimation for sentences?

4.2 A Preliminary Experiment

Which test set should we use to help answer our research questions? Before we report on our main set of experiments, we report on a preliminary set of experiments. Given the similarity of our task with the TREC QA track, it may be seem obvious to try and use the evaluation resources created there. However, when we tried to use TREC data for assessing our system, we faced some problems which forced us to set up separate assessments for our purposes. In this section, we briefly describe the result of a preliminary experiment carried out on TREC data and the associated problems.

From the 2004 TREC QA track we took 10 TREC QA topics that are about persons. As in the TREC QA setting, we use the AQUAINT corpus to search for target sentences. We focused on the so-called “other” questions since they closely match our needs. We used the nuggets and the associated sentences (assessed as “vital”) generated by TREC’s human assessors for the “other” questions as our gold standard. We ran the methods described in Section 3 on these 10 topics. All the methods returned a ranked list of sentences. We took the top 50 sentences of the output of these methods and assessed the results. Assessment is simple; we check each sentence in the output if it is found in the gold standard. We then compute the score based on the formula given in [22]. The result is summarized in Table 1.

Methods	F-Score
Baseline (M0)	0.548
Word overlap (M1)	0.560
Language modeling (M2)	0.564
Graph-based, generic (M3)	0.487
Graph-based, without term weights (M4)	0.567
Graph-based, with term weights (M5)	0.503

Table 1: Results from the preliminary experiment

Although the size of the data limits the scope of the claims that can be made on the basis of the results, the simpler methods seem to perform better. However, when we look at the outputs of the different methods, we immediately notice that some of the sentences ranked high by our methods contain important biographical information not included in TREC's ground truth, as shown by the following examples:

- **Fred Durst**: Born in Jacksonville, Fla., Durst grew up in Gastonia, N.C., where his love of hip-hop music and break dancing made him an outcast.

- **Eileen Marie Collins**: She was born Nov. 19, 1956, in Elmira, N.Y., to Jim and Rose Collins.

In the TREC setting, such sentences are often not included in the set of snippets deemed important—sometimes because they may be covered by corresponding factoid questions for the same scenario. As the goal of the current work is to develop methods for extracting any important biographical information, we do not make distinction between the different types of biographical facts (factoid vs lists vs "other"). As a result, we found it necessary to create separate assessments of the outputs of our systems, at least for the top ranking sentences. In this way, we believe, we can be sure that the assessments are "in sync" with our task definition. In the following, we resume our description of the experimental setup.

4.3 Experimental Set-up

Task Given a topic t (of category c) the task is to return up to 20 important sentences about t.

Test Corpus We used the AQUAINT corpus as the corpus from which target sentences have to obtained [4]; the corpus is a heterogeneous collection of over 1M news articles from multiple sources.

Test Topics As to our test topics, we restricted ourselves to people related topics, and randomly selected 10 occupations ("categories") and for each occupation we randomly selected 3 people with that occupation and at least one occurrence in Wikipedia (to be able to create reference corpora, see below). The categories and names are listed in Table 2, columns 1 and 2, respectively.

Ground truth For each method the output was manually assessed by two assessors, who followed a procedure similar to those followed by the TREC QA assessors [23]. That is, each assessor examined each of the sentences in the output and determined whether they are about the topic; if not, they were marked as non-relevant. If a sentence was marked relevant, the assessors determined whether it contains important (biographical) information. Assessors were allowed to examine the topic in Wikipedia or using a general purpose web search engine. Cohen's kappa was 0.70, which indicates substantial agreement [5]. In case of disagreement, the assessors worked together to come to a shared decision.

Evaluation measure As creating a recall base for the task of identifying important biographical facts is beyond our resources, we employed an early precision oriented evaluation measure. We proceeded as follows. If a sentence is judged to contain important information, it is marked as "V" (vital). If a sentence contains important information that is already contained in sentences higher up in the ranked list, it is marked as "VD" (vital duplicate). Systems are then compared based on the counts for "V" and the sum of "'V" and "VD" sentences.

Significance testing To determine whether the observed differences are statistically significant, we use the two-tailed Wilcoxon Matched-pair Signed-Ranks Test, and look for improvements at a significance level of 0.05 (*) and 0.01 (**).

Reference Corpora Methods M2, M3, M4, M5 rely on reference corpora for their importance estimation. To create reference corpora, we use Wikipedia. Most Wikipedia pages provide a relatively complete description of a given entity: they are authoritative fact files for their entities. We used pattern matching to extract Wikipedia category pages and select people related categories. For every people related category in Wikipedia—i.e., for every subcategory of *People*[2] (and all of their subcategories, etc)—we took a random sample of n Wikipedia pages labeled with that subcategory. Thus, for every such subcategory (and the corresponding occupation) we created a (small) corpus consisting of the n (n=30) randomly selected Wikipedia pages.

The reference corpora were equipped with a simple graph structure. According to Wikipedia's authoring guidelines, pages need to be structured in such a way that the *lead* (containing the leading paragraph(s) of an article) provides essential information about the topic of the page, effectively serving as a summary of the page.[3] Correspondingly, we incorporate a bias in favor of target sentences that are similar to reference sentences appearing close to the top of a Wikipedia page by organizing the reference sentences in a hierarchical graph: the first sentence receives a link from the 2nd, and the 2nd from the 3rd, etc.

4.4 Results

Table 2 shows the results of our scoring formula (1) using the six importance estimation methods described above. In the table, *WD* refers to the number of sentences that our assessors found to be important biographical descriptions counting duplicates as separate instances, whereas *WOD* indicates the number of important biographical description after removing duplicates. In the table below we summarize the significant differences; each method above the horizontal line significantly outperforms the methods listed below it in the same column.

M0	M1	M2	M3	M4	M5
M3**	M3**	M3*	–	M0**, M1**, M2**, M3**	M0**, M1**, M2**, M3**

When we look at the relative performance on the different topics, we see that some topics are "easy" for all methods, some are hard for all, and others display a mixed behavior. E.g., *Jack Welch* and *Kim Clijsters* are easy for all, while

[2] See `http://en.wikipedia.org/w/index.php?title=Category:People`

[3] See `http://en.wikipedia.org/wiki/Wikipedia:Guide_to_layout` for general guidelines.

Table 2: Test topics (column 2), with their category (column 1), and number of important sentences identified using methods M0–M5. For all algorithms, we display the number of important sentences returned in the top 20, both with duplicates (columns 4, 6, 8, 10, 12, 14), and without (columns 3, 5, 7, 9, 11, 13).

				Not graph-based				Graph-based					
		M0		M1		M2		M3		M4		M5	
Category	Name	WOD	WD	WOD	WD	WOD	WD	WOD	WD	WOD	WD	WOD	WD
Actress	Jean Harlow	6	8	8	12	8	8	4	4	8	11	8	15
	Charlize Theron	4	5	8	12	6	15	4	7	5	12	5	12
	Joan Collins	5	8	4	4	2	2	4	6	7	8	9	11
Actor	James Dean	5	5	3	3	4	4	5	5	6	6	6	6
	Broderick Crawford	2	3	1	1	0	0	1	2	2	2	1	1
	Glenn Ford	1	1	1	1	3	3	1	2	5	5	5	5
Architect	Frank Gehry	8	10	8	13	7	11	7	15	9	10	9	9
	Richard Rogers	1	2	2	5	1	1	1	1	5	10	5	9
	Gustave Eiffle	8	11	1	1	3	3	2	2	4	4	5	5
Astronaut	Eileen Marie Collins	8	11	3	18	6	16	2	20	5	15	7	14
	Boris Morukov	2	7	2	6	1	1	0	0	2	6	2	6
	Gennady Padalka	2	15	2	11	3	8	2	12	3	8	4	9
Boxer	Floyd Patterson	9	13	7	12	5	7	5	6	8	15	7	13
	Nigel Benn	6	13	6	9	0	0	4	9	10	16	9	14
	Laila Ali	3	4	3	15	6	9	4	11	6	12	5	11
Business Leaders	Jack Welch	9	10	7	17	6	12	6	11	11	14	10	13
	John Bogle	8	12	4	14	7	10	5	16	6	10	5	11
	Ted Turner	4	7	12	15	10	11	0	0	7	11	7	11
Musician	Fred Durst	6	8	5	10	2	6	3	18	8	15	9	14
	Bryan Adams	4	5	5	6	2	2	4	6	7	10	6	8
	Ice Cube	6	6	4	10	5	14	1	1	7	14	8	13
Novelist	Franz Kafka	5	7	5	8	7	13	5	5	6	9	7	8
	Muriel Spark	6	7	1	2	2	4	0	0	6	7	5	6
	Barbara Kingsolver	4	9	5	14	6	16	5	16	9	15	7	15
State Leader	Bashar Assad	11	12	12	17	7	15	4	14	9	10	11	14
	Frederick Chiluba	4	11	6	15	6	16	3	20	7	14	7	13
	Ion Iliescu	7	14	2	6	3	14	4	19	4	10	6	10
Tennis Player	Jennifer Capriati	3	7	9	15	17	20	4	15	11	15	13	14
	Martina Hingis	2	2	6	13	9	13	3	4	9	15	10	14
	Kim Clijsters	7	10	7	17	8	10	10	15	11	13	11	14
	Total	156	243	149	302	152	264	99	252	203	322	209	318

Broderick Crawford—an actor who lived from 1911 to 1986 and got an academy award in 1949—is hard for all, which is probably due to the fact that the topic is less prominent currently. More interestingly, on topics such as *Jennifer Cappriati*, *Nigel Benn*, and *Ted Turner*, we see that the scores of the various methods diverge considerably.

Returning to the questions raised at the outset of the section, we see that the use of a reference corpus actually hurts (but not significantly): the baseline outperforms M1 and M2. However, on top of the generic graph-based method (M3) it makes a significant difference (cf. M4 and M5). Furthermore, graph-based importance estimation does not necessarily improve over the baseline (cf. M3 vs M0), but it does when combined with a reference corpus (cf. M4 and M5 vs the baseline); indeed, M4 and M5 outperform the other methods on nearly all topics. Finally, while the explicit representation of important lexical items does make a positive difference, the difference is not significant (cf. M4 vs M5): M5 outperforms M4 on 11 topics, M4 outperforms M5 on 9 topics, while there is a draw on the remaining 10 topics.

4.5 A Closer Look

We examined the output of the systems; this revealed interesting qualitative differences between the results of the four systems. The baseline, M1, and M2 tend to return large numbers of sentences containing duplicate information. For example, most of the top ranking target sentences for *Eileen Marie Collins* mention that *she was the first woman to command a space shuttle*, whereas the top ranking sentences for *Jennifer Capriati* describe the results of tennis matches with several people and are of the type *Jennifer Capriati, United States, bt Kim Clijsters, Belgium, 7-5, 6-3*. Often, important facts are ranked very highly by M4 and M5, while other methods rank them lower (if at all).

Even though M1 and M2 do make use of reference corpora, they only seem to be able to identify “popular” sentences with respect to the newspaper corpus. In contrast, by coupling reference corpora to PageRank-style global importance estimation, M4 and M5 go beyond mere popularity. Further examination of the output of method M5 provides supporting evidence. In method M5, the middle layer, labeled “Words,” contains all the words that are shared by both the target and reference sentences and it serves as a bridge between the two sets of sentences. An interesting side-effect of the M5 method is that the shared lexical items are also ranked (along with the target sentences). Table 3 shows the top ten terms generated for two occupations (*Architect* and *Astronaut*) using this mechanism.

These terms can be taken as lexical features that characterize the respective occupations. Since these terms receive higher weights, sentences containing these

Architects	Astronauts
architect	space
left	mission
first	nasa
work	astronaut
architecture	flight
designed	first
design	born
most	named
known	served
years	became
buildings	pilot
life	years
building	selected

Table 3: Ranked lists of terms for *Architecture* and *Astronauts* as produced by the M5 method.

lexical items are likely to be ranked higher as exemplified by the top ranked sentence for the architect *Richard Rogers*.

- British **architect** Richard Rogers, best **known** for his **design** of the Pompidou Center in Paris and the Millennium Dome in London, got the award for **architecture**. (Richard Rogers)

Finally, the methods also differ on the textual unit used to compute the importance score. Methods M1, M4 and M5 use sentence-level (i.e., sentences in *reference corpus*) comparisons in computing the importance score whereas M2 is based on corpus-level (i.e., likelihood computed on the entire *reference corpus*) comparison in computing the importance score. Although M1 performs the same as M2 in terms of *WOD* counts, it is much closer to M4 and M5 in terms of *WD* counts (cf. Table 2). However, M1 bases its sentence selection decision on localized information, i.e., based on a single reference sentence (cf. Figure 1). Therefore, there is a tendecy for this method to select redundant information. On the other hand, M4 and M5 capture the global nature of importance estimation by using a graph-based method to combine sentence-level evidence. In general, this may suggest that methods that compute importance scores in a bottom-up manner by aggregating the evidence obtained at the sentence level tend to perform better. This, however, needs to be further investigated.

5 Related Work

Related work comes in several flavors: question answering, summarization, and novelty checking.

Question Answering Question Answering (QA) has attracted a great deal of attention, especially since the launch of the QA track at TREC in 1999. While significant progress has been made in technology for answering general factoids (e.g., *How fast does a cheetah run?*), there is a real need to go beyond such factoids [21]. At the TREC QA track this has been recognized through the introduction of definition questions and of so-called "other" questions (that ask for important information about a topic at hand that the user does not know enough about to ask). These "other" questions are part of so-called scenarios, while our "Tell me about X." questions are asked in isolation. For reasons explained in Section 2 we decided to develop our own test, following much of the TREC assessment guidelines. Similar scenarios are being examined at the WiQA pilot mentioned in Section 2, which has a lot in common with our task, although we view a Wikipedia only as a source from which to obtain reference corpora, not as as the source of target sentences.

One of the tasks at the 2004 edition of the Document Understanding Conference [9] was to provide a short summary which is relevant to the question "Who is X?", i.e., to provide a short biographical summary about the person X. Such questions were to be answered by returning important snippets that help define the person for whom a definition is being sought.

There are various strategies for answering definition questions and "other" questions. Some systems implement pattern matching techniques for identifying potential answers, based on either surface or linguistic structures [8, 13, 16, 19]. Others rely on knowledge bases built through offline mining of corpora, again based on surface patterns (such as [12]), or deeper linguistic analyses for extracting facts from a corpus [15]. One of our importance estimation methods (M1) is similar to a method introduced by [1], who used Wikipedia as "an importance model" in answering "other" questions. For a given topic, they extract candidate snippets from the AQUAINT corpus, and if the topic has an entry in Wikipedia, the snippets from the AQUAINT corpus are then ranked based on word overlap with the text in the Wikipedia entry. More recently, machine learning approaches are being explored by some [3, 6]. Unlike these approaches, our methods (including M4 and M5) are based on relatively shallow (generic) techniques and can easily be tested and deployed in other domains.

Summarization The graph-based part of methods M3, M4, M5 is related to ideas from extractive summarization, particularly graph-based summarization techniques and feature-based biographical information extraction [11]. [10] (LexRank) and [17] (TextRank) introduced a graph-based method for ranking sentences based on their relevance to document summaries. Though the method as originally proposed is not appropriate for the task of identifying important descriptions, a modification of the method, combined with reference sources such as Wikipedia, can be usefully and effectively applied for our task. This usage of Wikipedia is similar to ideas discussed in [11], but instead of explicitly generating highly specific occupation related lexical features, we use Wikipedia to generate reference corpora and leave the topic-related knowledge implicit.

Novelty checking Within the setting of the TREC Novelty track [20], Allan et al. [2] reviewed a number of methods for identifying relevant and novel sentences for a topic, which are mainly based on inter-sentence similarity measures. They showed that novelty detection largely depends on the quality of relevant sentence identification step. Though the ideas introduced there are directly relevant for the current work, our emphasis is on the estimation of importance which differentiates it from the more restricted novelty checking task.

6 Conclusion

We have presented a transparent schema for importance ranking that combines relevancy, aboutness, and an importance feature. Our focus was on methods for estimating importance features. Our main contribution is a combination of a corpus-based approach to capturing the knowledge encoded in sentences known to be important with a graph-based method for ranking sentences. Experimental evaluations show that this combination significantly improves over a number of competitive baselines.

Our best-performing methods require an auxiliary system that accurately predicts a category of a given entity. We are currently working on methods for doing just this from a sample description. Furthermore, we want to study the impact of the size of the corpora that capture knowledge about important facts. How large should the corpora be for our algorithm to be effective? How homogeneous should they be? Also, our overall sentence ranking schema has two other components besides an importance feature: “relevancy” and “aboutness.” We also plan to investigate the impact of these components.

Furthermore, we plan to experiment with more complex queries and more noisy data sets (instead of the AQUAINT collection).

Finally, in the context of Wikipedia, the method can be used to identify important sentences from an online source while editing or creating Wikipedia pages. We are investigating the use of our methods within the setting of the WiQA task at CLEF [14].

7 Acknowledgements

We are grateful to our referees for providing valuable feedback.

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001. In addition, De Rijke was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.069.006, 612.069.006, 640.-001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

[1] D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach. Using Wikipedia at the TREC QA Track. In *Proceedings TREC 2004*, 2004.

[2] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *In SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, 2003*, pages 314–321, 2003.

[3] I. Androutsopoulos and D. Galanis. A practically unsupervised learning method to identify single-snippet answers to definition questions on the web. In *HLT/EMNLP*, pages 323–330, 2005.

[4] AQUAINT. The AQUAINT Corpus of English News Text, 2002. URL: `http://www.ldc.upenn.edu/Catalog/docs/LDC2002T31/`.

[5] R. Bakeman and J. Gottman. *Observing interaction: An introduction to sequential analysis*. Cambridge University Press: New York, 1986.

[6] R. Barzilay and L. Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL 2004: Proceedings of the Main Conference*, pages 113–120, 2004.

[7] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[8] T. Clifton and W. Teahan. Bangor at trec 2004: Question answering track. In *Proceedings TREC 2004*, 2004.

[9] DUC. Document Understanding Conference, 2005. URL: `http://www-nlpir.nist.gov/projects/duc/`.

[10] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479, 2004.

[11] E. Filatova and J. Prager. Tell me what you do and i'll tell you what you are: Learning occupation-related activities for biographies. In *HLT-NAACL*, pages 49–56, 2004.

[12] M. Fleischman, E. Hovy, and A. Echihabi. Offline strategies for online question answering: answering questions before they are asked. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 1–7. Association for Computational Linguistics, 2003.

[13] W. Hildebrandt, B. Katz, and J. Lin. Answering definition questions with multiple knowledge sources. In *HLT-NAACL*, pages 49–56, 2004.

[14] V. Jijkoun and M. de Rijke. Overview of WiQA 2006. In A. Nardi, C. Peters, and J. Vicedo, editors, *Working Notes CLEF 2006*, September 2006.

[15] V. Jijkoun, M. de Rijke, and J. Mur. Information extraction for question answering: Improving recall through syntactic patterns. In *Proceedings of the 20th International on Computational Linguistics (COLING 2004)*, 2004.

[16] H.-J. Lee, H.-J. Kim, and M.-G. Jang. Descriptive question answering in encyclopedia. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 21–24, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[17] R. Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings ACL 2004*, 2004.

[18] D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th intern. conf. on World Wide Web*, pages 13–19, New York, NY, USA, 2004. ACM Press.

[19] B. Schiffman, I. Mani, and K. J. Concepcion. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *ACL*, pages 450–457, 2001.

[20] I. Soboroff and D. Harman. Novelty Detection: The TREC Experience. In *Proceedings HLT/EMNLP 2005*, pages 105–112, 2005.

[21] R. Soricut and E. Brill. Automatic question answering: Beyond the factoid. In *Proceedings HLT/NAACL*, 2004.

[22] E. Voorhees. Overview of the TREC 2004 question answering track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, 2005. NIST Special Publication: SP 500-261.

[23] E. Voorhees and H. Dang. Overview of the trec 2005 question answering track. In *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*, 2005.

[24] J. Xu. TREC2003 QA at BBN: answering definitional questions. In *Proceedings TREC 2003*, 2003.