A Method for Short Message Contextualization: Experiments at CLEF/INEX

Liana Ermakova

A Method for Short Message Contextualization: Experiments at CLEF/INEX

Liana Ermakova*

Institut de Recherche en Informatique de Toulouse, Toulouse, France
Perm State National Research University, Perm, Russia
ermakova@irit.fr

Abstract. This paper presents the approach we developed for automatic multi-document summarization applied to short message contextualization, in particular to tweet contextualization. The proposed method is based on named entity recognition, part-of-speech weighting and sentence quality measuring. In contrast to previous research, we introduced an algorithm from smoothing from the local context. Our approach exploits topic-comment structure of a text. Moreover, we developed a graph-based algorithm for sentence reordering. The method has been evaluated at INEX/CLEF tweet contextualization track. We provide the evaluation results over the 4 years of the track. The method was also adapted to snippet retrieval and query expansion. The evaluation results indicate good performance of the approach.

Keywords: Information retrieval, tweet contextualization, summarization, snippet, sentence extraction, readability, topic-comment structure

1 Introduction

The efficient communication tends to follow the principle of the least effort. According to this principle, using a given language interlocutors do not want to work any harder than necessary to reach understanding. This fact led to the extreme compression of texts especially in electronic communication, e.g. microblogs, SMS, search queries. However, sometimes these texts are not self-contained and need to be explained since understanding of them requires knowledge of terminology, named entities (NE) or related facts. The idea to contextualize short texts like micro-blogs or tweets is quite recent. Meij et al. mapped a tweet into a set of Wikipedia articles but in their work, no summary is provided to the user, rather a set of related links [6]. San Juan et al. went a step further and introduced Tweet Contextualization (TC) as an INEX task which became the CLEF lab in 2012 [9, 2].

The main motivation of this research is to help a user to better understand a short message by extracting a context from an external source like the Web or the Wikipedia by means of text summarization. A summary is either an "extract", if it consists in the most important passages extracted from the original

* Ambassade de France en Russie, bourse de thèse en cotutelle

text, or an "abstract", if these sentences are re-written, generating a new text. In this paper we focus on extracts. Extraction implies two steps: (1) searching for relevant sentences and (2) organizing them into a readable text. In previous summarization approaches sentence retrieval is based on the similarity to the query [10]. We also use this principle. In addition, we assume that part-of-speech (POS) tagging can ameliorate results since in general some POS provide more information than others (e.g. nouns are more informative that adverbs or functional words). As in [5], we integrated POS weights into the TF-IDF measure. The application of NE recognition may improve information retrieval (IR) performance, including tweet study [8], therefore we introduced NE similarity measure. Not all sentences are suitable for summarization purpose (e.g. headers, labels etc.). To avoid trash passages we enriched our method by sentence quality measure based on Flesch reading ease test, lexical diversity, meaningful word ratio and punctuation ratio. Thus, the proposed approach is based on NE recognition, POS weighting and sentence quality measuring.

Usually, a sentence is viewed as a unit in summarization task. However, often a single sentence is not sufficient to catch its meaning and even human beings need a context. In contrast to [13], we believe that a context does not provide redundant information, but allows to precise and extend sentence meaning. Therefore, we introduce an algorithm to smooth a candidate sentence by its local context, i.e. the neighboring sentences from the source document. Neighboring sentences influence the sentence of interest, but this influence decreases as the remoteness of the context increases, which differs from the previous approaches where the dependence is considered to be binary (i.e. a neighboring sentence influences the sentence of interest or not) [7]. The binary understanding of the influence of the context assumes that the influence is the same for all sentences.

Moreover, our algorithm takes advantage of topic-comment structure of sentences. The topic-comment structure have already got the attention of linguists in the 19-th century, however, it is hardly applied in IR tasks. To our knowledge, the topic-comment analysis was never exploited in the summarization task.

As Barzilay et al. showed, sentence order is crucial for readability [1]. Moreover, sentence reordering is the only way to improve the readability of a text produced by an extraction system. Barzilay et al. proposed to order the sentences by searching for the Hamiltonian path of maximal length in a directed graph where vertices are themes and edges corresponds to the number of times a theme precedes the other one. This approach requires a training corpus. In contrast to this, we hypothesized that in a coherent text neighboring sentences should be somehow similar to each other and the total distance between them should be minimal. Therefore, we propose an approach to increase global coherence of text on the basis of its graph model, where the vertices correspond to the extracted passages and the edges represent the similarity measure between them. Under these assumptions, sentence ordering implies searching for the minimal path that visits each vertex exactly once. This task is known as the traveling salesman problem. However, this method does not consider chronological constraints therefore we introduce another method based on the sequential ordering

problem. In contrast to [1], our approach is not restricted by the news articles on the same topic and it takes advantages of the similarity between sentences.

The proposed approach demonstrated better performance than other systems like Cortex, Enertex, REG, etc. Cortex combines such metrics as word frequency, overlap with query terms, entropy of the words, shape of text etc. [11]. In Enertex sentence score is calculated from text energy matrix [11]. REG is an enhancement of Cortex which uses query expansion (QE) [12].

The rest of the paper is organized as follows. Section 2 presents our method. Section 3 contains the results and their analysis. Section 4 suggests the application of the proposed sentence retrieval method to snippet generation and QE. Section 5 concludes the paper.

2 Method Description

We participated in the INEX TC Track that aims at evaluating systems providing a context to a tweet. A context should be a readable summary of a limited size (up to 500 words) extracted from the Wikipedia dump. In this section we present our approach and its evolution over four-year period. The proposed method aims at contextualizing short messages by extracting passages from an external text collection. In this case contextualization task can be considered as query-biased multi-document summarization where a short message corresponds to a query. Our approach includes three steps: (1) preprocessing of the queries and the corresponding documents; (2) sentence scoring; and (3) sentence re-ordering.

Query preprocessing involves hashtag and reply treatment as well as combining different query parts. We put higher weight to words occurring in hashtags. We split hashtags and replies by capitalized letters. An initial tweet is expanded by the words obtained from tweet hashtags and replies as stated above. Thus, a tweet *RT StateDept: #SecKerry: Europe is strong, and stronger together. Europe and the US together have an opportunity to create jobs, build a stronger future* is expanded by *State, Dept, Sec, Kerry*. We assume that relevant sentences come from relevant documents. Documents are retrieved by the Terrier platform[1]. We apply a DFR (divergence from randomness) model InL2c1.0 which is a default retrieval model in Terrier based on TF-IDF measure with L2 term frequency normalization. 5 top-ranked documents are considered. Queries and documents are parsed by Stanford CoreNLP[2] which integrates such tools as POS tagger and NE recognizer. Parser annotation is merged with Wikipedia tags.

2.1 Sentence Scoring

In 2011 we introduced a system based on TF-IDF cosine similarity measure, special weighting for POS, NE, structural elements of a document, definitional sentences and the algorithm for smoothing from local context. Prior scores of

[1] terrier.org/
[2] nlp.stanford.edu/software/corenlp.shtml

sentence r_i was a product of the cosine similarity measure sim_{uni} between the sentence and the query that included IDF and POS weight and the NE similarity sim_{NE}:

$$r_i = sim_{uni} \times sim_{NE} \tag{1}$$

$$sim_{NE} = \frac{NE_{common} + NE_{weight}}{NE_{query} + 1} \tag{2}$$

where NE_{common} is the number of NE appearing in both query and sentence, NE_{query} is the number of NE appearing in the query, NE_{weight} is positive floating point parameter that allows not to reject sentence without NE which can be still relevant. We add 1 to the denominator to avoid division by zero.

We introduced an algorithm for smoothing from the local context. We assumed that the neighboring sentences influence the sentence of interest, but this influence decreases as the remoteness of the context increases. In other words, the nearest sentences should produce more effect on the target sentence sense than others. We choose the simplest dependence model, namely the linear function. In this case, the smoothed relevance $R(S)$ is calculated by the formulas:

$$R(S) = \sum_{i=-k}^{k} w_i \times r_i, \ \sum_{i=-k}^{k} w_i = 1 \tag{3}$$

$$w_i = \begin{cases} \frac{1-w(S)}{k+1} \times \frac{k-|i|}{k} & 0 < |i| \leq k \\ w(S), & i = 0 \end{cases} \tag{4}$$

where $w(S)$ is the weight of the sentence S set by a user, w_i and r_i are respectively the weights and the prior scores of the sentences from the context of S of k length. If the sentence number in left or right context is less than k, their weights are added to the target sentence weight $w(S)$. This allows keeping the sum equal to one since otherwise a sentence with a small number of neighbors (e.g. the first or last sentences) would be penalized.

In 2011 our system showed the best results according the relevance judgment (see [3] for details). In 2012 we modified our method by adding bigram similarity, anaphora resolution, hashtag processing, redundancy treatment and sentence reordering. However, we obtained lower results than in the previous year. Therefore, in 2013 we decided to not consider bigram similarity, anaphora resolution, nor redundancy treatment. We also used generalized POS (e.g. we merge regular adverbs, superlative and comparative into a single adverb group). To avoid trash passages we enriched our method by sentence quality measure based on Flesch reading ease test, lexical diversity, meaningful word ratio and punctuation ratio. Lexical diversity allows avoiding sentences that do not contain terms except those from the query. We define it as the number of different lemmas used within a sentence divided by the total number of tokens in this sentence. Meaningful word ratio over the total number of tokens in the sentence is aimed at penalizing sentences that either have no sense at all or are not comprehensible without large context. The punctuation score penalizes sentences containing

many punctuation marks. Thus, we believe that a good sentence should have high ratio of different meaningful words and reasonable ratio of punctuation.

The sentence score $score(S)$ is estimated as the product of its quality $Q(S)$, smoothed relevance $R(S)$ and the score of the document $DocRel(d)$ from which it is extracted:

$$score(S) = DocRel(d) \times Q(S) \times R(S) \tag{5}$$

We define sentence quality $Q(S)$ as the product of the lexical diversity $Div(S)$, Flesch index $F(S)$, meaningful word ratio $M(S)$ and punctuation score $P(S)$:

$$Q(S) = Div(S) \times M(S) \times P(S) \times F(S) \tag{6}$$

$$P(S) = 1 - \frac{PM(S)}{T(S)} \tag{7}$$

where $PM(S)$ is the number of punctuation marks in S, and $T(S)$ is the number of tokens in S. $P(S)$ shows the ratio of tokens that are not punctuation marks.

2.2 Topic-comment Relationship in Contextualization Task

Linguistics establishes the difference between the clause-level topic and the discourse-level topic. The discourse-level topic refers to the notion of aboutness. While most IR models make the assumption that relevant documents are about the query and that aboutness can be captured considering bags of words only, we rather consider a clause-level topic-comment structure. The topic (or theme) is the phrase in a clause that the rest of the clause is understood to be about, and the comment (also called rheme or focus) is what is being said about the topic. In most languages the common means to mark topic-comment structure are word order, intonation and special constructions. In simple English clause the topic usually coincides with the subject. Therefore, topic identification in our approach is performed under assumption of topic fronting, i.e. the tendency to place topic at the beginning of a clause. We simplify this hypothesis by assuming that topic should be place at the sentence beginning. Sentence beginning is viewed as the first half of the sentence.

In 2014 participants should provide a context to tweets from the perspective of the related entities. Tweets are at least 80 characters long and do not contain URLs. A tweet has the following annotation types: the category (4 distinct), an entity name from the Wikipedia (64 distinct) and a manual topic label (235 distinct) (see an example Table 1). The context has to explain the relationship between a tweet and an entity. As in previous years it should be a summary extracted from a Wikipedia dump. We hypothesize that topic-comment relation-

Table 1. Tweet example 2014

tweet_id	category	entity	topic	content
213051315880869888	automotive	Fiat	sales	Seeing a lot of #Fiat cars downtown these days. #Traffic

ship identification is useful for this task. Quick query analysis provides evidence that an entity may be considered as a topic, while tweet content refers rather to comment, i.e. what is said about the entity. In order to link an entity to a tweet we combined the fields entity, topic and content into a single search query. Moreover, we assumed that providing the context to an entity implies that this context should be about the entity, i.e. the entity is the topic, while the retrieved context presents the comment. We used these assumptions for candidate sentence scoring. We doubled the weight of sentences in which the topic contains the entity under consideration.

2.3 Sentence Re-ordering

Although sentence ordering was not evaluated at INEX, we propose an approach to increase global coherence of text based on its graph model. The hypothesis is that neighboring sentences should be somehow similar to each other and the total distance between them should be minimal since word repetition is one of the formal indicators of text coherence. In our approach vertices represent sentences and edges correspond to the distances between adjacent sentences estimated as $1 - sim_{uni}$. If two relevant sentences are neighbors in the original text, they are considered as a single vertex. Thus, we reduced sentence ordering task to traveling salesman problem (TSP). TSP is an NP-hard problem in combinatorial optimization. Given a graph, the task is to find the shortest path that visits each vertex exactly once and returns to the start vertex. Algorithms to find the exact solution have exponential complexity. Therefore, we chose the greedy nearest neighbor algorithm with minor changes. Since sentence ordering does not request to return to the start vertex and the start vertex is arbitrary, we tried every vertex as the start one and chose the best result, i.e. the start vertex giving the path of the minimal length.

However, this method does not consider chronological constraints. Sentences with time stamps (e.g. date and time) should be ordered chronologically. Other sentences are not restricted by the chronological constraints but the coherence of text should be the maximal. As in the TSP approach, we believe that text coherence increases as the total sum of the distances between neighboring sentences decreases, i.e. the similarity between adjacent sentences should be maximal. So, we modified the task and it gave us sequential ordering problem (SOP). SOP "is a version of the asymmetric traveling salesman problem where precedence constraints on the vertices must also be observed" [4]. SOP is stated as follows. Given a directed graph, find a Hamiltonian path of the minimal length from the start vertex to the terminal vertex observing precedence constraints. Usually SOP is solved by the means of integer programming. Integer programming is NP-hard and these methods achieved only limited success. Therefore, we solved the problem as follows. Firstly, we ordered sentences with time stamps assigned by a parser $s_1 - s_2 - ... - s_n$. Sentences without time stamp were added to the set $P = \{p_j\}_{j=\overline{1,m}}$. For each pair $s_i - s_{i+1}$ we searched for the shortest path passing through vertices from P. These vertices were removed from P and $i = i + 1$.

If $i = n$, we searched for the shortest path passing through the vertices that remained in P and the edge with the maximal weight was removed.

3 Evaluation

In this paper we focus on the results demonstrated at INEX in the two last years. Summaries were evaluated according to their informativeness and readability.

Informativeness was estimated as the lexical overlap (*uni, big* and *skip* representing the proportion of shared unigrams, bigrams and bigrams with gaps of two tokens respectively) of a summary with the pool of relevant passages extracted from the runs submitted by all participants [2]. Official ranking was based on decreasing score of divergence with the gold standard estimated by *skip*:

$$Dis(S,T) = \sum_{t \in T} \frac{f_{T(t)}}{f_T} \times \left(1 - \frac{\min \log P, \log Q}{\max \log P, \log Q}\right) \quad (8)$$

where $P = \frac{f_{T(t)}}{f_T} + 1$ and $Q = \frac{f_{S(t)}}{f_S} + 1$, T is the set of terms in the pool of relevant passages, $f_{T(t)}$ is the frequency of a term t (*uni, big* or *skip*) in the pool, $f_{S(t)}$ is the frequency of a term t in a summary.

In 2013 the informativeness was estimated as the overlap of a summary with 3 pools of relevant passages: (1) prior set (PRIOR) of relevant pages selected by organizers (40 tweets, 380 passages); (2) pool selection (POOL) of the most relevant passages (1 760) from participant submissions for 45 selected tweets; and (3) all relevant texts (ALL) merged together with extra passages from a random pool of 10 tweets (70 tweets, 2 378 relevant passages) [2]. The system was evaluated with three parameter sets. In our run 273 each sentence is smoothed by its local context and first sentences from Wikipedia article which it is taken from. The run 274 has the same parameters except it does not have any smoothing. In our best run 275 punctuation score is not taken into account, it has slightly different formula for NE comparison and no penalization for numbers. Among automatic runs our best run 275 was ranked first (PRIOR and POOL) and second (ALL) over 24 runs submitted by all participants. Table 2 provides results of the best automatic systems presented by the participants. Our results are marked by *. The best results are set off in bold. According to bigrams and skip bigrams, our best run is 275, while according to unigrams the best run is 273. So, we can conclude that smoothing improves Informativeness. Another conclusion is that ranking is sensitive to the pool selection as well as to the choice of divergence.

In 2014 there were 240 tweets in English collected by the organizers of CLEF RepLab 2013. 2 gold standards (1/5 of the topics) were used: (1) pool of relevant sentences per topic (SENT); and (2) pool of noun phrases (NOUN) extracted from these sentences together with the corresponding Wikipedia entry. The first run (ETC) was performed by the system developed in 2013. Three fields (entity, topic and content) were treated as a query. An entity was treated as a single phrase. The second run (ENT) differed from ETC by double weight for sentences

Table 2. Informativeness evaluation 2013

Run	All.skip	All.big	All.uni	Pool.skip	Pool.big	Pool.uni	Prior.skip	Prior.big	Prior.uni
258	**0,894**	**0,891**	**0,794**	0,880	0,877	0,792	0,929	0,923	0,799
275*	0,897	0,892	0,806	**0,879**	**0,875**	0,794	**0,917**	**0,911**	0,790
273*	0,897	0,892	0,800	0,880	0,875	**0,792**	0,924	0,916	**0,786**
274*	0,897	0,892	0,801	0,881	0,875	0,793	0,923	0,915	0,787

where the entity represented the topic. The third run (RESTR) was based on document set retrieved for the tweet and filtered by the results obtained for the entity. Thus, the document retrieved by using the field content as a query were rejected if they did not coincide with top-ranked documents retrieved by using the field entity. According to the evaluation performed on the pool of sentences, our runs ETC, ENT and RESTR were ranked 3-rd, 4-nd and 6-th; while according to the evaluation based on noun phrases, they got slightly better ranks, namely 2, 3 and 5 respectively. Thus, the best results among our runs were obtained by the system that merges fields entity, topic and content into a single query. The run #360 is better than our runs according to sentence evaluation; nevertheless, it showed worse results according to noun phrase evaluation. Our system is targeted at nouns and especially NEs. This could provoke the differences in ranking with respect to sentences and noun phrases. The run based on entity restriction showed worst results. This could be explained by the fact that filtering out the documents that are considered irrelevant to the entity may cause a big loss of relevant documents if they are not top-ranked according to entities. The results of ETC and ENT are very close. However, topic-subject identification slightly decreased the performance of the system. Yet we believe that finer topic-comment identification procedure may ameliorate the results.

Table 3. Informativeness evaluation 2014

Run	SENT.uni	SENT.big	SENT.skip	NOUN.uni	NOUN.big	NOUN.skip
361	**0.7632**	**0.8689**	**0.8702**	**0.7903**	**0.9273**	**0.9461**
360	0.782	0.8925	0.8934	0.8104	0.9406	0.9553
ETC^*	0.8112	0.9066	0.9082	0.8088	0.9322	0.9486
ENT^*	0.814	0.9098	0.9114	0.809	0.9326	0.9489
$RESTR^*$	0.8152	0.9137	0.9154	0.8131	0.936	0.9513

Readability was estimated as mean average (MA) scores per summary over relevancy (T), soundness (no unresolved anaphora) (A), non-redundancy (R) and syntactical correctness (S) among relevant passages of the ten tweets having the largest text references. The score of a summary was the average normalized number of words in valid passages. Sentence order was not judged at INEX/CLEF.

In 2013 according to all metrics except redundancy our approach was the best among all participants (see Table 4). Runs were officially ranked according to mean average scores. Readability evaluation also showed that the run 275 is the best by relevance, soundness and syntax. However, the run 274 is much better in terms of avoiding redundant information. The runs 273 and 274 are close according readability assessment as well.

In 2014 we received very low score for diversity and structure. This may be related to the fact that we decide not to treat this problem since in previous years their impact was small. Despite we retrieved the entire sentences from the Wikipedia, unexpectedly we received quite low score for syntactical correctness.

Table 4. Readability evaluation 2013

Rank	Run	MA	T	R	A	S
1	275	**72.44**%	**76.64**%	67.30%	**74.52**%	**75.50**%
2	274	71.71%	74.66%	**68.84**%	71.78%	74.50%
3	273	71.35%	75.52%	67.88%	71.20%	74.96%

4 Other Applications of the Sentence Retrieval

Our approach is generic enough to be applied for various tasks. Here, we consider two of them: snippet retrieval and query expansion.

4.1 Snippet Retrieval

A search engine returns a larger number of results that a user cannot examine all. Therefore, a search engine provides a user with snippets (small text passages appearing under a search result extracted from the document) to help in evaluating web page relevance before browsing it.We slightly modified the method applied for TC for the INEX Snippet Retrieval Track 2012-2013: (1) nominal sentences were not penalized; (2) sentences were not re-ordered; (3) we did not treat redundancy since in the single-document summarization the probability of redundant information is much lower, and snippets are short and should be generated fast. We used two algorithms for the candidate passage selection: dynamic programming approach to solve the knapsack problem and the moving window (MW) algorithm.

A snippet is limited up to 1-2 sentences (150-300 symbols) but it should provide as much information about the underlying document as possible. Therefore, snippet retrieval can be viewed as a task of selecting passages of the maximal total importance under the restriction of the total weight. This task is known as a knapsack problem stated as follow: given a set of items (sentences), each with a weight (number of symbols) and a value (score), find the subset of this set to pack the rucksack so that the total weight is less than or equal to a given capacity and the total value is as large as possible. We solve this problem by the basic dynamic programming algorithm $DP - 1$.

However, this algorithm has pseudo-polynomial time. Moreover, if each sentence within a document were greater than a predefined threshold, the snippet would be an empty string. Therefore, we used a MW algorithm to find the best scored passage. At each step the first token is removed from a candidate passage and the tokens following the candidate passage are added while its total weight is no greater than a predefined threshold. The passage with the maximal score is selected as a snippet. Despite the most relevant information may occur in the too long sentences, snippets beginning in the middle of a sentence have lower readability. That is why, we penalize them.

Evaluation was performed manually by the organizers of INEX Snippet Retrieval Track 2013 [2]. The relevance of the documents was judged apart from the relevance of the snippets. Then these judgments were integrated by the following measures: Mean prediction accuracy (MPA), Mean normalized prediction accuracy (MNPA), Recall, Negative recall (NR), Positive agreement (PA), Negative agreement (NA), and Geometric mean (GM). The official ranking was based on GM. The results are given in the Table 5 (our results are marked by *, the best values are set off in bold). Our approach demonstrated the highest performance. As we hypothesized, the knapsack algorithm provided better results since it searches for the most valuable information regardless its position.

Table 5. Snippet evaluation 2013

Rank	Run	MPA	MNPA	Recall	NR	PA	NA	GM
1	*knapsack**	**0.8300**	**0.6834**	**0.4190**	0.9477	**0.4921**	0.8673	**0.5352**
2	Focused	0.8171	0.6603	0.3507	0.9700	0.4210	0.8675	0.4774
3	Focused_Split	0.8214	0.6549	0.3684	0.9413	0.4358	0.8624	0.4732
4	MW^*	0.8300	0.6459	0.3852	0.9067	0.4283	0.8572	0.4605
5	Baseline	0.8171	0.6414	0.2864	**0.9964**	0.3622	**0.8711**	0.4025

4.2 Query Expansion

QE in a search engine may be also viewed as contextualization of the initial query. The key idea of the proposed method is to search the most appropriate

candidates for QE by ranking terms and sentences from the pseudo-relevance feedback. Our approach is underlain by the following hypotheses: (1) good expansion terms come from quality sentences relevant to the query; (2) they should have appropriate POS and high IDF; and (3) the terms lying in the neighborhood of query terms are closer related to them than the remote ones. Candidate terms are ranked according to the following metric:

$$w_{total}(t) = score(S) \times w_{pos}(t) \times IDF(t) \times importance(t, Q) \tag{9}$$

$$importance(t, Q) = wd(t, Q) \times cooccurrence(t, Q) \tag{10}$$

where $score(S)$ is score of the sentence S containing t computed by (5), $w_{pos}(t)$ is the weight of the POS of t, $IDF(t)$ is the inverse document frequency of the candidate term, $wd(t, Q)$ is a function of the distance from the candidate terms to the query Q and their weights, and $coocurence(t, Q)$ shows the likelihood of the candidate term to occur not by chance with the query terms in the top documents ranked according to the initial query. Our approach outperformed the baseline InL2c1.0 and DFR models for QE (KL, CS, Bo1, Bo2) implemented in Terrier according to MAP, NDCG, R-precision, P@5, P@10, and P@100 on TREC Ad Hoc 6-8 collection and WT10g. The differences between the our approach and other evaluated methods are significant at the level $p < 0.05$ for TREC Ad Hoc 6-8. On WT10g the differences with Bo2 and KL models are not significant.

5 Conclusion

In this paper we presented an approach for short message contextualization from an external source based on query-biased summarization. Our approach implies sentence retrieval and re-ordering. Sentence retrieval is based on NE recognition, POS weighting and sentence quality measuring. We introduced an algorithm of smoothing from the local context. We also integrated the knowledge of topic-comment structure into the sentence retrieval model. Moreover, we developed a graph-based algorithm for sentence re-ordering. The method has been evaluated at INEX/CLEF TC track. We obtained the best results in 2011 according to informative evaluation. In 2013 according to informative evaluation our system was ranked first (PRIOR and POOL) and second (ALL) over all automatic systems that participated. At the same time in terms of readability it was the best among all participants according to all metrics except redundancy. Run comparison showed that smoothing improves informativeness. Another conclusion is that ranking is sensitive to the pool selection as well as to the choice of divergence. Despite the topic-comment analysis did not improve results, we believe that small changes in implementation may produce positive effect on the system performance. In 2014 the worst results among our runs were shown by the run based on entity restriction that could be explained by the loss of the recall. Although sentence ordering was not evaluated at INEX campaign, we believe that it is crucial for readability. The sentence retrieval method was also adapted to snippet retrieval and QE. In 2013 our system showed the best results in the INEX Snippet Retrieval Track.

Although sentence ordering was not evaluated at INEX, we propose an approach to increase global coherence of text based on its graph model. The hypothesis is that neighboring sentences should be somehow similar to each other and the total distance between them should be minimal since word repetition is one of the formal indicators of text coherence. In our approach vertices represent sentences and edges correspond to the distances between adjacent sentences estimated as $1 - sim_{uni}$. If two relevant sentences are neighbors in the original text, they are considered as a single vertex. Thus, we reduced sentence ordering task to traveling salesman problem (TSP). TSP is an NP-hard problem in combinatorial optimization. Given a graph, the task is to find the shortest path that visits each vertex exactly once and returns to the start vertex. Algorithms to find the exact solution have exponential complexity. Therefore, we chose the greedy nearest neighbor algorithm with minor changes. Since sentence ordering does not request to return to the start vertex and the start vertex is arbitrary, we tried every vertex as the start one and chose the best result, i.e. the start vertex giving the path of the minimal length.

However, this method does not consider chronological constraints. Sentences with time stamps (e.g. date and time) should be ordered chronologically. Other sentences are not restricted by the chronological constraints but the coherence of text should be the maximal. As in the TSP approach, we believe that text coherence increases as the total sum of the distances between neighboring sentences decreases, i.e. the similarity between adjacent sentences should be maximal. So, we modified the task and it gave us sequential ordering problem (SOP). SOP "is a version of the asymmetric traveling salesman problem where precedence constraints on the vertices must also be observed" [4]. SOP is stated as follows. Given a directed graph, find a Hamiltonian path of the minimal length from the start vertex to the terminal vertex observing precedence constraints. Usually SOP is solved by the means of integer programming. Integer programming is NP-hard and these methods achieved only limited success. Therefore, we solved the problem as follows. Firstly, we ordered sentences with time stamps assigned by a parser $s_1 - s_2 - ... - s_n$. Sentences without time stamp were added to the set $P = \{p_j\}_{j=\overline{1,m}}$. For each pair $s_i - s_{i+1}$ we searched for the shortest path passing through vertices from P. These vertices were removed from P and $i = i + 1$. If $i = n$, we searched for the shortest path passing through the vertices that remained in P and the edge with the maximal weight was removed.

In this chapter we proposed a novel approach to document re-ranking in information retrieval based on topic-comment structure of texts. Although it can be easily generalized to document retrieval. To the best of our knowledge, this information structure was never applied to the ad hoc information retrieval nor re-ranking. We introduced an automatic topic-comment annotation method based on the topic fronting assumption that requires only shallow parsing, namely sentence chunking and POS tagging. The main idea of the proposed method is to split a sentence into two parts by a personal verb. We integrated topic-comment structure into BM25F retrieval model. Firstly, we hypothesized that the topics should have more weight than the comments. However, the preliminary studies

demonstrated that high values of this coefficient decreased the results in average. The possible explanation is that the comments are usually much longer than the topics and therefore the prior probability of a query term to occur within comments is higher. Higher values of topic weight could lead to the lost of documents that just mention relevant information but are not entirely about the subject. We evaluated our approach on two TREC data sets. According to all used evaluation measures for both test collections, our method significantly outperformed the strong baseline provided by the Terrier platform. Experiment results allow drawing a conclusion that the approach proposed in this chapter is more suitable for difficult queries. Since our method makes the difference between sentences where the topic and the comment are inversed (as in **??** and **??**), we believe that our approach makes sense for question answering and focused IR. In future work we are going to investigate these tracks.

References

1. Barzilay, R., Elhadad, N., McKeown, K.R.: Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research pp. 35–55 (2002), 17
2. Bellot, P., Doucet, A., Geva, S., Gurajada, S., Kamps, J., Kazai, G., Koolen, M., Mishra, A., Moriceau, V., Mothe, J., Preminger, M., SanJuan, E., Schenkel, R., Tannier, X., Theobald, M., Trappett, M., Wang, Q.: Overview of INEX 2013. In: Forner, P., Mller, H., Paredes, R., Rosso, P., Stein, B. (eds.) Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Lecture Notes in Computer Science, vol. 8138, pp. 269–281. Springer Berlin Heidelberg (2013)
3. Ermakova, L., Mothe, J.: IRIT at INEX: Question answering task. In: Geva, S., Kamps, J., Schenkel, R. (eds.) Focused Retrieval of Content and Structure, Lecture Notes in Computer Science, vol. 7424, pp. 219–226. Springer Berlin Heidelberg (2012)
4. Herndvlgyi, I.T.: Solving the sequential ordering problem with automatically generated lower bounds. Proceedings of Operations Research 2003 pp. 355–362 (2003)
5. Lioma, C., Blanco, R.: Part of speech based term weighting for information retrieval. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 5478, pp. 412–423. Springer Berlin Heidelberg (2009)
6. Meij, E., Weerkamp, W., de Rijke, M.: Adding semantics to microblog posts. In: Proceedings of the fifth ACM international conference on Web search and data mining. p. 563572. WSDM '12, ACM, New York, NY, USA (2012)
7. Murdock, V.G.: Aspects of sentence retrieval. Dissertation (2006)
8. de Oliveira, D.M., Laender, A.H., Veloso, A., da Silva, A.S.: FS-NER: A lightweight filter-stream approach to named entity recognition on twitter data. In: Proceedings of the 22Nd International Conference onArabic named entity recognition World Wide Web Companion. pp. 597–604. WWW '13 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2013)
9. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J.: Overview of the INEX 2011 question answering track (QA@INEX). In: Geva, S., Kamps, J., Schenkel, R. (eds.) Focused Retrieval of Content and Structure, Lecture Notes in Computer Science, vol. 7424, pp. 188–206. Springer Berlin Heidelberg (2012)

10. Shen, C., Li, T.: Learning to rank for query-focused multi-document summarization. pp. 626–634. IEEE (2012)
11. Torres-Moreno, J.M., Velzquez-Morales, P., Gagnon, M.: Statistical summarization at QA@INEX 2011 track using cortex and enertex systems. In: Geva, S., Kamps, J., Schenkel, R. (eds.) Focused Retrieval of Content and Structure, Lecture Notes in Computer Science, vol. 7424, pp. 247–256. Springer Berlin Heidelberg (2012)
12. Vivaldi, J., da Cunha, I.: QA@INEX track 2011: Question expansion and reformulation using the reg summarization system. In: Geva, S., Kamps, J., Schenkel, R. (eds.) Focused Retrieval of Content and Structure, Lecture Notes in Computer Science, vol. 7424, pp. 257–268. Springer Berlin Heidelberg (2012)
13. Yang, Z., Cai, K., Tang, J., Zhang, L., Su, Z., Li, J.: Social context summarization. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 255–264. ACM, Beijing, China (2011)