# DCU@INEX-2012: Exploring Sentence Retrieval for Tweet Contextualization

Debasis Ganguly, Johannes Leveling, and Gareth J. F. Jones

CNGL, School of Computing, Dublin City University, Dublin 9, Ireland
{dganguly, jleveling, gjones}@computing.dcu.ie

**Abstract.** For the participation of Dublin City University (DCU) in the INEX-2012 tweet contextualization task, we investigated sentence retrieval methodologies. The task requires providing the context to an ad-hoc real-life tweet. This context is to be constructed from Wikipedia articles. Our approach involves indexing the passages in Wikipedia articles as separate retrievable units, extracting sentences from the top ranked passages, computing the sentence selection score for each such sentence with respect to the query, and then returning the top most similar ones. The simple sentence selection strategy performed quite well in the task. Our best run has ranked first from the *readability* perspective and ranked eighth as ordered by *informativeness* out of 33 official runs.

## 1 Introduction

The tweet contextualization task was first introduced at INEX in 2011. The task requires construction of a short summary so as to explain the context associated with a given tweet. This context information has to be constructed from Wikipedia articles. As an example, for the CNN tweet "RT @CNNLive: View stake-out camera at funeral home where #WhitneyHouston body is expected to arrive in New Jersey. Watch live: `http://t.co/nyqT4PUa`", the system is expected to provide such expository information as who is Whitney Houston, what is she famous for, how did she die etc.

The task being different from standard ad-hoc IR poses with its own set of challenges. Firstly, the tweet text is very different from keyword based queries of ad-hoc search or web search. This necessiatates applying pre-processing steps on the tweet texts to get an appropriate query string. For example, the tweet hash-tags do not exist in Wikipedia articles and needs to be appropriately processed to get a useful query term. Secondly, a standard passage retrieval may not be suitable for the task because of the restriction on the length in the reported summary. The text in a passage itself may surpass the length theshold requirement of the summary. It thus makes sense to decompose passages into smaller units, i.e. sentences, and collate them together.

Previous approaches to INEX-QA have mostly used passage retrieval coupled with a summarizer. Sentence retrieval on the other hand has widely been employed in TREC-QA tasks for both factoid and definition question answering [1].

Sentence retrieval has the potential to perform well for tweet contextualization because sentences being short contain more focussed information than the relatively larger passages which may contain digressory content. Furthermore, the effect of sentence retrieval on tweet contextualization has still been unexplored. This motivated us to apply various sentence retrieval strategies on the tweet contextualization task.

## 2 System Description

In this section, we describe our system details. After describing the document and query processing, and retrieval, we focus on to our working methodologies for sentence selection.

### 2.1 Document Indexing

We used a modified version of the SMART[1] system for the experiments at INEX 2012. Each paragraph from the Wikipedia corpus[2] was indexed as a retrievable document unit. The beginning of a passage is marked by the XML tag <*p*>. This resulted in a total of over 26M passages to retrieve from. Extracted portions of documents, namely text under the <*title*>, <*p*>, <*h*>, <*t*> tags, were indexed using single terms and a controlled vocabulary (or pre-defined set) of statistical phrases following Salton's blueprint for automatic indexing [2]. Stopwords that occur in the standard stop-word list included within SMART were removed. Words were stemmed using a variation of the Lovins' stemmer implemented within SMART. Frequently occurring word bi-grams (loosely referred to as phrases) were also used as indexing units. We used the N-gram Statistics Package (NSP)[3] on the English Wikipedia text corpus from INEX 2006 and selected the 100,000 most frequent word bi-grams as the list of candidate phrases.

### 2.2 Query Processing

The tweet texts were pre-processed to produce queries to retrieve against the indexed collection. The pre-processing steps are described as follows. The URLs from the tweets were removed employing a regular expression based pattern matcher. *Medial capital* words, i.e. words with inner uppercase letters, were split into separate words e.g. the word "WhitneyHouston" was decomposed into "Whitney" and "Houston". Tweet hash-tags were split up into the prefix # character followed by the word, e.g. "#Whitney" was decomposed into # and "Whitney".

The following word breaking rules were applied to split hashtags starting with "#" and usernames starting with "@": A break between the last and current character is employed if:

[1] `ftp://ftp.cs.cornell.edu/pub/smart/`

[2] `http://dev.termwatch.es/esj/Term2IR/2012/data/tweetcontext2012corpus.xml.gz`

[3] `http://www.d.umn.edu/~tpederse/nsp.html`

i) the last character is lower case and the current character is upper case or digit (e.g. "OccupyWallStreet" ->"Occupy Wall Street");
ii) the last character is upper case and the last character of a valid acronym, the current character is also upper case or a digit (e.g. "CNNNews" -> "CNN News";
iii) the last character and the current character have different case and the resulting word would be longer than 3 characters.

### 2.3 Retrieval

The context for each tweet was constructed in two passes as follows. In the first pass, we retrieved $N$ passages using language modelling (LM) [3] similarity with Jelinek-Mercer smoothing. The smoothing parameter $\lambda$ was set to 0.6. In the second pass, we score senteneces based on three different methodologies, explained later in details. We then concatenate the top $M$ sentences until the length of the concatenated summary string exceeds the threshold of 500 characters limit. The concatenation step ensures that we do not add duplicate sentences in the summary.

### 2.4 Sentence Retrieval Methodologies

**Language Modelling Similarity.** The most simple sentence scoring technique is that of scoring a sentence $S$ by its LM score computed with respect to the query i.e. the pre-processed tweet text. This is done as shown in Equation 1.

$$P(S|Q) \propto \prod_{q \in Q} \mu P(q|S) + (1-\mu)P(q) \tag{1}$$

Note that the smooting parameter $\mu$ used in Equation 1 is different from $\lambda$ which was used for retrieving the passages as discussed in Section 2.3.

**Relevance Model Similarity.** The second sentence selection strategy which we use is derived from relevance model (RLM) term scores [4]. The key idea in RLM-based retrieval is that relevant documents and query terms are assumed to be sampled from an underlying hypothetical model of relevance R pertaining to the information need expressed in the query. In the absence of training data for the relevant set of documents, the only observable variables are the query terms and the top-ranked $R$ pseudo-relevant documents assumed to be generated from the relevance model. Thus, the estimation of the probability of a word $w$ being generated from the relevance model is approximated by the conditional probability of observing $w$ given the observed query terms. Thus higher a word $w$ co-occurs with a query tern $q$, higher is the likelihood of $w$ to be sampled from the relevance model, i.e. higher is $P(w|\mathrm{R})$. This is shown in Equation 2.

$$P(w|q_i) \propto \sum_{j=1}^{R} P(w|D_j)P(q_i|D_j) \tag{2}$$

We can easily extend this notion of relevance model weighting of terms to whole sentences by simply aggregating over the constituent words of a sentence. This is shown in Equation 3 which we use to score every sentence and select the top-scoring ones in the returned summary.

$$P(S|\mathrm{R}) = \prod_{w \in S} P(w|\mathrm{R}) \tag{3}$$

**Topical Relevance Model Similarity.** This sentence selection score is based on an extended version of relevance model (RLM) similarity. In our extended relevance model, we compute the probabilities $P(w|D)$s by marginalizing them over a set of latent topics. Firstly, we estimate the topic distribution over the set of top ranked passages retrieved in the initial step by latent Dirichlet allocation (LDA) [5]. LDA outputs two distribution vectors $\theta$ (from document to topic) and $\phi$ (from topic to word). Modified smoothed document models are obtained by using these two distributions as shown in Equation 4, where $K$ is the number of topics used in the LDA estimation.

$$P(w|D) = \sum_{k=1}^{K} P(w|z_k, \phi) P(z_k|D, \theta) \tag{4}$$

We then use the topic smoothed document models in the estimation of RLM i.e. we use the definition of $P(w|D)$ as obtained from Equation 4 in 2 to obtain an extended RLM sentence selection methodology which we name topical relevance model (TRLM) similarity.

$$P(w|q_i) \propto \sum_{j=1}^{R} \big( \sum_{k=1}^{K} P(w|z_k, \phi) P(z_k|D_j, \theta) \big) P(q_i|D_j) \tag{5}$$

## 3 Run Description

We submitted three official runs (run ids: 185, 186 and 187) for the INEX-2012 Tweet contextualization task. The first pass passage retrieval for each of the three runs is identical and follows the description of Sections 2.1, 2.2 and 2.3. The sentence retrieval strategies of each of these runs is different. Run 185 used simple language modelling (LM) similarity, run 186 used RLM similarity, whereas run 187 used TRLM similarity to score sentences. The number of top documents used for the (T)RLM estimation was set to 20. For TRLM, the additional parameter $K$, i.e. the number of topics, was set to 5. Our submissions did not use any automatic summarization techniques for sentence selection. We rather relied on pure IR-based approaches to generate the twweet contexts.

## 4 Evaluation

The tweet contexts were evaluated with two measures: a) *informativeness*, which measures the *closeness* of the answer string with a golden reference with the

**Table 1.** Official results for INEX-2012 Tweet contextualization task

| Run Id | Run Description | Rank | Informativeness Metrics | | |
|---|---|---|---|---|---|
| | | | Uni-gram | Bi-gram | Skip-gram |
| 185 | LM sentence retrieval | 8 | 0.8265 | 0.9129 | 0.9135 |
| 186 | RLM sentence retrieval | 10 | 0.8347 | 0.9210 | 0.9208 |
| 187 | TRLM sentence retrieval | 11 | 0.8360 | 0.9235 | 0.9237 |
| 178 | Official best | 1 | **0.7734** | **0.8616** | **0.8623** |
| 194 | Organizers' baseline | 4 | 0.7864 | 0.8868 | 0.8887 |

| Run Id | Run Description | Rank | Readability Metrics | | |
|---|---|---|---|---|---|
| | | | Relevance | Syntax | Structure |
| 185 | LM sentence retrieval | 1 | **0.7728** | **0.7452** | **0.6446** |
| 186 | RLM sentence retrieval | 5 | 0.7008 | 0.6676 | 0.5636 |
| 187 | TRLM sentence retrieval | 14 | 0.6093 | 0.5252 | 0.4847 |
| 194 | Organizers' baseline | 4 | 0.6975 | 0.6342 | 0.5703 |

help of KL divergence between the two; and b) *readability*, which measures the syntactic coherence of the text such as whether it has grammatical errors, has unresolved anaphora or is redundant etc [6].

Table 1 reports the official results of our three submitted runs. Along with our runs, the table shows the offcial best run as measured by informativeness and also the run submitted by the organizers as the baseline. Informativeness evaluation involves computation of three metrics: the KL divergence between the golden summary and the returned summary for uni-grams, bi-grams, and bi-grams with two allowable gaps in between [6]. Note that KL divergence being a distance measure implies that a lower value of this metric is indicative of a better result. The readability metric on the other hand reports the proportion of text which has correct syntax, structure and is relevant in the context. As a result, a higher value of these metrics indicates a better result.

It can be seen that the most simple sentence retrieval technique using LM similarity fairly well, achieving rank eight, as measured by informativeness. This run in fact achieves the best readability result.

Our other runs, i.e. the (T)RLM based sentence selection strategies, have not performed well in the official evaluation. The release of official relevance assessments namely the reference summary context for each tweet would enable us to tune the parameters of the two other sentence selection strategies in order to achieve an improved performance.

## 5 Conclusions and Future work

In our first participation at the INEX Tweet contextualization task, we applied sentence retrieval to construct answer fragments for each tweet. Three different

sentence selection methodologies were used: i) language modelling (LM) score, ii) relevance modelling (RLM) scoring of a sentence by accumulating over the RLM scores of its constituent terms, and iii) topical relevance modelling (TRLM) scoring of a sentence by accumulating over the topic smoothed RLM scores of its constituent terms.

The results confirm that simple IR-based sentence selection techniques can perform fairly well on both the informativeness and the readability metrics, without the application of any complex NLP techniques. The main advantage of the sentence retrieval methodologies is that these are very fast in contrast to computationally intensive NLP methods.

## Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.

## References

1. Voorhees, E.M.: Overview of the TREC 2003 question answering track. (2003) 54–68
2. Salton, G.: A Blueprint for Automatic Indexing. ACM SIGIR Forum **16**(2) (Fall 1981) 22–38
3. Hiemstra, D.: Using Language Models for Information Retrieval. PhD thesis, Center of Telematics and Information Technology, AE Enschede (2000)
4. Lavrenko, V., Croft, B.W.: Relevance based language models. In: Proceedings of the SIGIR '01, ACM (2001) 120–127
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3** (2003) 993–1022
6. SanJuan, E., Moriceau, V., Tannier, X., Bellot, P., Mothe, J.: Overview of the INEX 2011 Question Answering Track (QA@INEX). In: Pre-proceedings of the INitiative for the Evaluation of XML retrieval workshop (INEX 2011), Saarbrcken (Germany) (December 2011) 145–153