The Effect Of Smoothing In Language Models For Novelty Detection

Ronald T. Fernández
Departamento de Electrónica y Computación.
Universidad de Santiago de Compostela
Campus Sur, s/n.
15782 Santiago de Compostela, SPAIN
ronald.teijeira@rai.usc.es

The novelty task consists of finding relevant and novel sentences in a ranking of documents given a query. In the literature, different techniques have been applied to address this problem. Nevertheless, little is known about Language Models for novelty detection and, especially, the effect of smoothing on the selection of novel sentences. Language Models can be used to study novelty and relevance in a principled way. These statistical models have been shown to perform well empirically in many Information Retrieval tasks. In this work we study formally the effects of smoothing on novelty detection. To this aim, we compare different techniques based on the Kullback-Leibler divergence and we analyze the sensitivity of retrieval performance to the smoothing parameters. The ability of Language Modeling estimation methods to handle quantitatively the uncertainty associated to the use of natural language is a powerful tool that can drive the future development of novelty-based mechanisms.

Keywords: Language Models, Smoothing, Novelty Detection, Kullback-Leibler Divergence.

1. INTRODUCTION

Novelty detection is an important research topic whose applications span a wide range of Information Retrieval (IR) problems [6]. We adopt here the novelty detection task as defined in the TREC conference [3, 8, 9]. The groups participating in this task start from a common ranking of documents for each query. The task is decomposed into two problems: 1) to produce a ranked set of relevant sentences (sentence retrieval stage), and 2) to filter out redundant sentences from this set (novelty detection stage). The task is an effort to go beyond the typical ranked list of documents. It explores retrieval techniques that return relevant and novel sentences (i.e. key sentences) rather than whole documents containing extraneous or duplicate information. Although the effectiveness of the first stage (sentence retrieval) is an important issue [1], we focus here on novelty detection. In particular, we are interested in Language Modeling (LM) techniques and the effect of smoothing on the selection of novel sentences.

LM is a principled statistical framework that has proved to work well in different areas, such as Speech Recognition, Machine Translation and Information Retrieval (IR). The original proposal to apply LM for IR was done by Ponte and Croft [7] and it was followed by a number of studies dedicated to the subject [4, 2]. Smoothing is a core problem in LM estimation. It adjusts the maximum likelihood estimator so as to correct the inaccuracy due to data sparseness. The type and level of smoothing affects directly performance in document retrieval [11]. However, the role of smoothing in novelty detection is largely unknown and there are not studies reporting how novelty techniques behave with varying levels of smoothing.

In this paper we evaluate novelty detection applying LMs estimated using different smoothing techniques. We study how the performance of novelty detection changes as the level of smoothing is increased. We compare aggregate models, where the set of seen sentences is modeled by a single LM, and non-aggregate models, where every seen sentence is handled by an individual LM. This study helps to gain insight into the role that LM estimation can play in current novelty detection systems. Most novelty techniques applied in the past (e.g. New Words, Set Difference or Cosine Distance [1]) are rather simplistic and lack the estimation power which is inherent to LMs. Moreover, many of the initial novelty approaches were only tested with the TREC 2002 novelty track collection. This is problematic because this collection contains little redundancy.

The rest of this paper is organized as follows. Section 2 reviews some papers related to our research. In section 3 we briefly explain the LM estimation of methods used and Kullback-Leibler divergence. The empirical evaluation conducted is reported in section 4. The paper ends with some conclusions.

2. RELATED WORK

The novelty detection methods experimented in the context of the TREC novelty tracks scan the output of a sentence retrieval component (a list of sentences ranked in decreasing order of similarity to a given query) and discard the sentences that do not contain new material. Initially, the ranked list of presumed relevant sentences is re-ordered in the order given by the task (sentences are considered in the same order in which the relevant

documents were originally ranked and multiple sentences from the same document are considered in the order in which they appear in the document). The first sentence is often assigned the highest score of novelty and the remaining sentences are scored in terms of some measure of overlapping between the sentence and the previously seen sentences. Simple word count measures, such as New Words, Set Difference or Cosine Distance [1] have been applied successfully in the past.

Some studies, such as the one conducted in [5], have applied LMs based on the Kullback-Leibler Divergence (KLD). KLD is applied to measure the divergence between a LM computed for a given sentence and a LM associated to the previously seen sentences. These LMs are maximum likelihood models smoothed using linear interpolation.

Another work [1] evaluates novelty using different smoothing techniques. The models based on Dirichlet Smoothing, Shrinkage Smoothing and Sentence Core Mixture Model apply pair-wise comparison between the LM of the current sentence and every LM of the seen sentences. The minimum divergence (computed applying KLD) is used to estimate the degree of novelty of the current sentence. The LMs are computed as follows. Shrinkage Smoothing and Sentence Core Mixture Model apply linear interpolation between three different LMs: a sentence LM, a topic LM and a general English LM. Dirichlet Smoothing simply applies smoothing over the sentence models in a (sentence) length-dependent way. The model based on Interpolate Aggregate Smoothing uses the KLD between the LM of the current sentence and a LM constructed from all previously seen sentences. These LMs are estimated using Jelinek-Mercer smoothing.

Nevertheless, there is no evidence about the performance of novelty detection when different levels of smoothing are applied. In our work we evaluate several smoothing techniques with different parameter settings and examine the sensitivity of novelty detection to the smoothing parameters. Our experimental setting is complete as we use the three TREC novelty track collections. This leads to a deep analysis on novelty performance under very different scenarios.

3. LANGUAGE MODELS

A Statistical Language Model is a probabilistic mechanism for explaining the generation of text. It basically defines a distribution over all possible word sequences. The simplest LM is the unigram LM, which is a word distribution. In this work we employ unigram LMs, whose effectiveness for IR tasks has been demonstrated in the literature [11].

A simple LM for a document *d* is the maximum likelihood estimator (*mle*). It associates a probability greater than zero for each term which appears in the document and a zero probability for the unseen terms. More specifically, for each term *w*, the probability $P_{mle}(w|d)$ represents the relative frequency of the term *w* in *d*.

This estimator is problematic because assigning probabilities equal to zero to any unseen term may be very strict. To overcome this problem, *mle*s are often smoothed using some fallback model that suffers less from sparseness (e.g. a model constructed from a large collection of documents). These smoothing techniques are explained in the next section.

3.1 Smoothing

Smoothing techniques try to balance the probability of terms which appear in a document with those ones which are missing. It discounts the probability mass assigned to the seen words and distributes the extra probability to the unseen terms according to some fallback model.

Jelinek-Mercer smoothing involves a linear interpolation of the maximum likelihood model with the collection model, using a coefficient λ.

$$P(w \mid d) = (1 - \lambda)P_{mle}(w \mid d) + \lambda P(w \mid C) = (1 - \lambda)\frac{c(w;d)}{\sum c(w;d)} + \lambda P(w \mid C)$$

where $P(w|C)$ is the *mle* constructed from the set of documents in the collection (*C*) and $c(w;d)$ the term count of *w* in *d*.

Dirichlet smoothing adjusts the amount of reliance on the observed text according to the length of this text:

$$P(w \mid d) = \frac{c(w;d) + \mu P(w \mid C)}{\sum c(w;d) + \mu}$$

where μ is the smoothing parameter.

As argued in [10], applying Dirichlet smoothing with query likelihood leads to a retrieval formula with components similar to the tf-idf weights and a document length correction.

3.2 KLD

KLD measures the divergence between two probability distributions. It can be used as a "distance" [*] between LMs. KLD is always positive and bigger than zero.

$$KLD(d_1 \,||\, d_2) = \sum_w P(w \mid d_1) \log \frac{P(w \mid d_1)}{P(w \mid d_2)}$$

The two smoothing techniques explained above and KLD have been jointly applied to model novelty. The subsequent experiments are reported in the next section.

4. EXPERIMENTS

We used the three collections of data which were made available in the context of the TREC Novelty tracks in 2002, 2003 and 2004 [3, 8, 9]. In 2002 and 2003 the ranking of documents given a query provided by NIST consists only of relevant documents. In 2004 the collection is more realistic because the ranked set of documents contains both relevant and non-relevant documents.

Given the ranked list of documents, the groups had to locate the sentences in the documents that are relevant and novel. Each document was also automatically split into sentences at NIST and sentences were assigned identifiers.

To study novelty detection properly we need first to rank sentences in decreasing order of presumed relevance (first stage: sentence retrieval). To get this initial ranking we applied a variation of tf-idf that has shown to be very effective and robust in the past [1]:

$$R(s|q) = \sum_{t \in \epsilon} \log(tf_{t,q}+1) \log \frac{n+1}{0.5+sf_t}$$

where $tf_{t,q}$ is the number of times term t occurs in the query, $tf_{t,s}$ is the number of times term t occurs in the sentence, sf_t is the number of sentences in which term t appears, and n is the number of sentences in the collection being scored.

Given this relevance ranking, we first experimented with two different baselines to detect novelty. The first baseline (named BNN – Baseline with No Novelty detection) ranks sentences using directly its relevance score. Sentences with higher tf-idf scores are placed in top positions in the ranking. This means that the novelty-oriented ranking of sentences is exactly the same as the relevance ranking (i.e. the highest the tf-idf similarity between the sentence and the query, the more novel the sentence is assumed to be). The second baseline (named BDOC – Baseline ordered by DOCument) consists of a reordering of the sentences from the relevance ranking. The sentences are considered in the same order in which the documents were originally ranked by NIST and multiple sentences from the same document are considered in the order in which they appear in the document. These baselines have been used in the past [1, 6] but there is not any comparative study which evaluates its relative merits for novelty detection.

In our experiments, BDOC performed clearly better than BNN. This is not surprising because the assessors that produced the sentence-level novelty judgments followed the same re-ordering policy taken in BDOC (i.e. they first identified the relevant sentences and, next, these sentences are considered in the order given by the task). We therefore set the BDOC baseline to be the reference baseline in our experiments.

We experimented with two different LM techniques for novelty detection. The first one, called NAM (Non-Aggregate Model) generates a smoothed LM for each sentence and computes the KLD between the LM of the current sentence s_i and the LM of every seen sentence s_j (where j=0,…,i-1). The minimum KLD between the sentence s_i and the seen sentences s_j is used as the novelty score for the sentence s_i. The second technique, the AM (Aggregate Model), generates only two LMs: a smoothed LM for the current sentence (s_i) and another LM for the set of seen sentences s_j (where j=0,…,i-1). The KLD between these two models is used to measure the degree of novelty of the sentence s_i.

We applied two different smoothing techniques to compute the LMs: Jelinek-Mercer and Dirichlet.

To study performance we computed precision at 10 sentences (P@10) and precision at 30 sentences (P@30). For the sake of brevity, we only analyze here the P@30 ratios (P@10 results showed similar trends but there was not significant differences between the baseline and the LM runs).

We applied these techniques for both short (title) and long (verbose) queries but trends were similar for both cases (we report here the results for short queries).

Figures 1 to 3 show results for NAM and AM applying Dirichlet and Jelinek-Mercer smoothing on the three collections.

[*] Note that it is not symmetric and does not satisfy the triangle inequality.

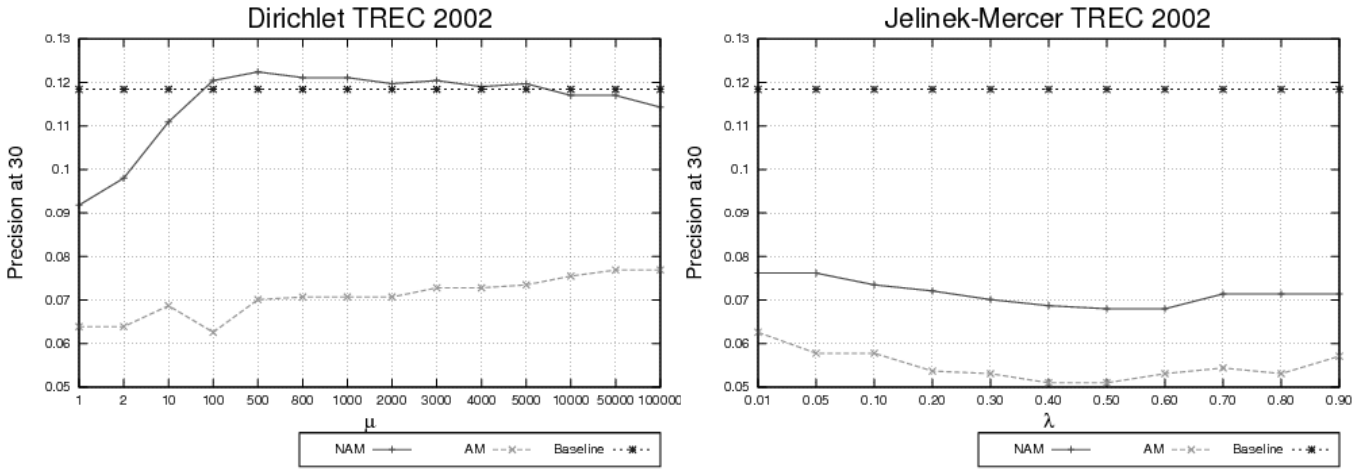

FIGURE 1: Dirichlet and Jelinek-Mercer Smoothing results applied over the TREC 2002.

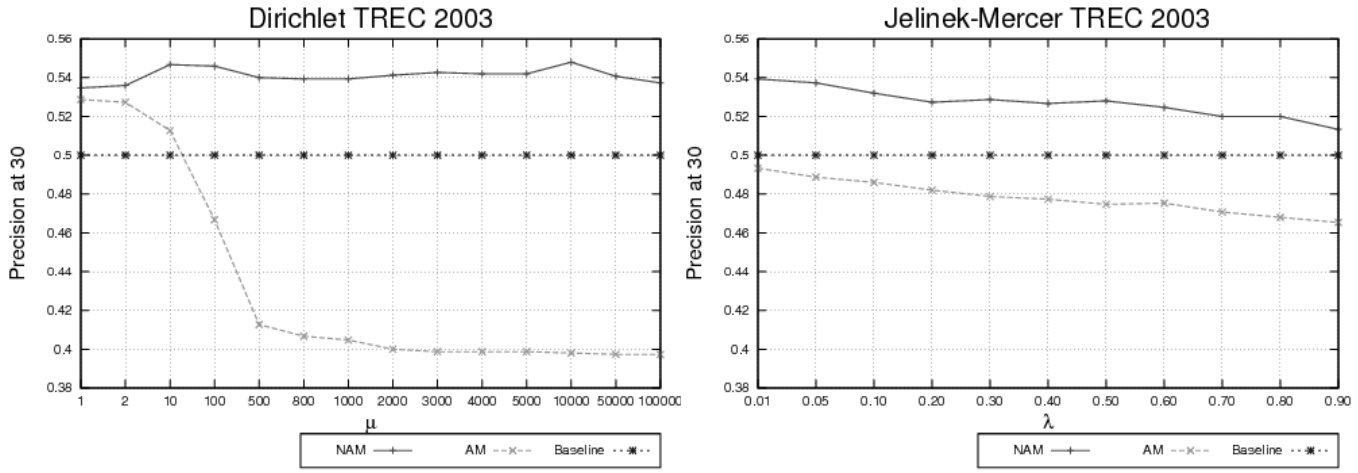

FIGURE 2: Dirichlet and Jelinek-Mercer Smoothing results applied over the TREC 2003.

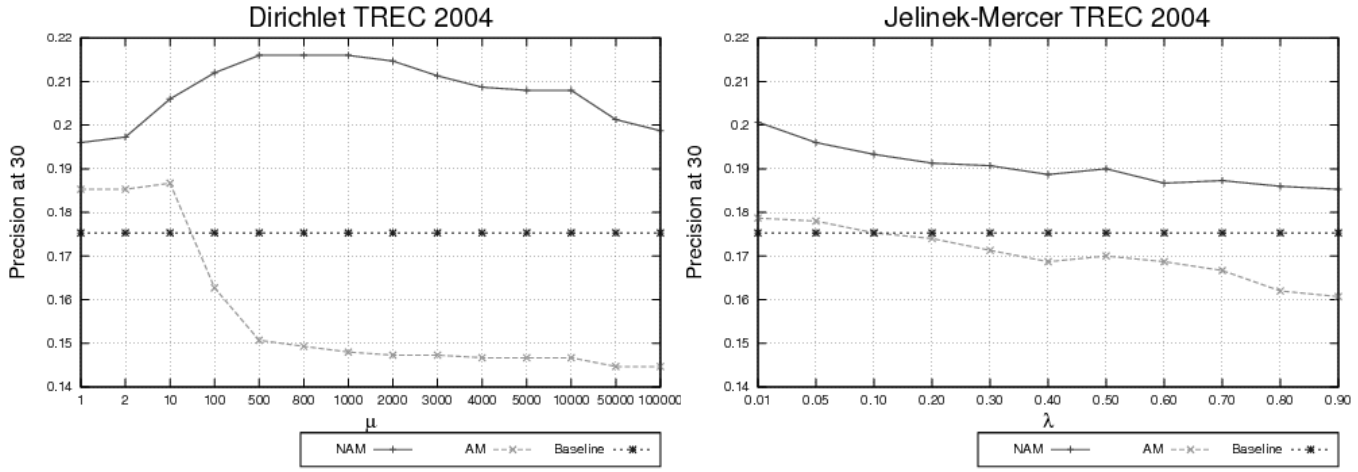

FIGURE 3: Jelinek-Mercer Smoothing results applied over the TREC 2004.

In TREC 2002, precision at 30 is significantly worse than in the other two collections; there are few relevant sentences in TREC (≈2%) and, therefore, the initial sentence retrieval stage works poorly. In fact, the results achieved by the groups participating in the TREC 2002 novelty track are similar to ours.

The experiments reported show some interesting trends. On one hand, Dirichlet smoothing performs better than Jelinek-Mercer. In document retrieval this was also the case as indicated in [11]. Dirichlet smoothing looks slightly

more sensitive to the parameter settings than Jelinek-Mercer but, still, nearly all Dirichlet runs improved over any Jelinek-Mercer run.
On the other hand, NAM performs always better than AM. NAM generates an individual LM for each sentence and, given a sentence s_i, its degree of novelty is estimated from the previously seen sentence having the smallest divergence. On the other hand, AM considers the history of seen sentences as a whole. This seems to be harming. A LM for the set of seen sentences might be too general. Consider a sentence which is an exact repetition of a past sentence. In NAM the sentence would receive the lowest novelty score. In AM, this is not guaranteed. The larger the history is, the less important the terms of the sentence are in the LM of the seen sentences. Therefore, it is still possible that the sentence is classified as novel. Note also that AM performs worse as smoothing increases. This is quite natural because, as we make the LM more general (we give more importance to terms in the collection), terms seen in the history receive increasingly less importance.
In TREC 2002 the baseline is very competitive and no LM run was able to improve over the baseline. This is naturally explained by the population of novel sentences in this collection. More than 90% of the relevant sentences were judged as novel by the TREC 2002 assessors. This means that a basic re-ordering of a relevance ranking (BDOC) is enough and no additional novelty oriented adjustments are needed. In contrast, the other two collections have much more redundancy and, therefore, the LM approaches improve over the baseline.
In all NAM cases, the optimum performance of Dirichlet smoothing tends to be found when the smoothing parameter (μ) is around 1000. The best precision with Jelinek-Mercer smoothing is reached when λ is small (0.01).
These results are preliminary and we still need to conduct further analysis on the different trends found. At the moment, AM does not seem a good technique to model the history of seen sentences because it can potentially "hide" the redundant pieces of texts into a global LM where many terms have non-marginal probabilities. In contrast, NAM looks quite effective but we still need to design new experiments and comparisons against other models.

5. CONCLUSIONS

In this paper we studied novelty detection at the sentence level using smoothed LMs and KLD. We focused our work in studying the performance of applying different smoothing techniques (Dirichlet smoothing and Jelinek-Mercer smoothing) with varying parameter setting.
To study novelty we applied two different techniques to build the LMs (NAM – Non-Aggregate Model – and AM – Aggregate Model) and used KLD to estimate the divergence between such models. We observed that NAM performs better than AM. We also showed that Dirichlet smoothing is a better estimation method than Jelinek-Mercer for novelty detection purposes. This observation agrees with the results shown in [11] for document retrieval. Dirichlet smoothing looks therefore a very robust smoothing method that works properly in different IR tasks. The comparison reported here is more exhaustive than other reports published in the literature because we have used the three TREC novelty collections available.

6. ACKNOWLEDGEMENTS

My thanks to my PhD. advisor, Dr. David E. Losada, who supervised this research. This work was partially supported by projects TIN2005-08521-C02-01 and PGIDIT06PXIC206023PN; and the Galician network 2006/23.

REFERENCES

[1] J. Allan, C. Wade, and A. Bolivar (2003). Retrieval and novelty detection at the sentence level. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), pages 314-321.
[2] W. B. Croft and J. Lafferty (Eds.) (2003). Language Modeling for Information Retrieval. Kluwer, 2003.
[3] D. Harman (2002). Overview of the TREC 2002 Novelty Track. Proceedings of the 11th Text Retrieval Conference (TREC 2002), Gaithersburg, MD.
[4] D. Hiemstra (2001). Using Language Models for Information Retrieval. Ph.D. Thesis, Centre for Telematics and Information Technology, University of Twente, January 2001.
[5] L. S. Larkey, J. Allan, M. E. Connell, A. Bolivar and C. Wade (2002). UMass at TREC 2002: Cross Language and Novelty Tracks.
[6] X. Li (2006). Sentence Level Information Patterns for Novelty Detection. Ph.D. Thesis, Department of Computer Science, University of Massachusetts at Amherst, September 2006.
[7] J. Ponte and W. B. Croft (2002). A language modeling approach to information retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998), pages 275-281, 1998.
[8] I. Soboroff (2004) Overview of the TREC 2004 Novelty Track. Proceedings of the 13th Text REtrieval Conference (TREC 2004) , Gaithersburg, MD.

[9] I. Soboroff and D. Harman (2003). Overview of the TREC 2003 Novelty Track. Proceedings of the 12th Text REtrieval Conference (TREC 2003) , Gaithersburg, MD.
[10] C. X. Zhai (2002). Risk Minimization and Language Modeling in Text Retrieval. Ph.D. Thesis, Language Technologies Institute, Carnegie Mellon University, July 2002.
[11] C. X. Zhai and J. Lafferty (2001). A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001), pages 334-342.