Título: IMPROVING SEARCH EFFECTIVENESS IN SENTENCE RETRIEVAL AND NOVELTY DETECTION

Nombre: Teijeira Fernandez, Ronald

Universidad: Universidad de Santiago de Compostela

Departamento: Electrónica y computación

Fecha de lectura: 20/05/2011

Mención a doctor europeo: concedido

Programa de doctorado: Interuniversitario en Tecnología de la Información

Dirección:

> **Director:** DAVID ENRIQUE LOSADA CARRIL

Tribunal:

> **presidente:** Senén Barro Ameneiro

> **secretario:** Alvaro Barreiro García

> **vocal:** Leif Azzopardi

> **vocal:** PABLO CASTELLS AZPILICUETA

> **vocal:** ENRIQUE ALFONSECA CUBERO

Descriptores:

> INTELIGENCIA ARTIFICIAL

> CIENCIA DE LOS ORDENADORES

El fichero de tesis ya ha sido incorporado al sistema

> 2011teijeimpro.pdf

Localización: BIBLIOTECA XERAL USC

Resumen: In this thesis we study thoroughly sentence retrieval and novelty detection. We analyze the strengths and weaknesses of current state of the art methods and, subsequently, new mechanisms to address sentence retrieval and novelty detection are proposed.

Retrieval and novelty detection are related tasks: usually, we initially apply a retrieval model that estimates properly the relevance of passages (e.g. sentences) and generates a ranking of passages sorted by their relevance. Next, this ranking is used as the input of a novelty detection module, which tries to filter out redundant passages in the ranking.

The estimation of relevance at sentence level is difficult. Standard methods used to estimate relevance are simply based on matching query and sentence terms. However, queries usually contain two or three terms and sentences are also short. Therefore, the matching between query and sentences is poor. In order to address this problem, we study in this thesis how to enrich this process with additional information: the context. The context

refers to the information provided by the surrounding sentences or the document where the sentence is located. Such context reduces ambiguity and supplies additional information not included in the sentence itself. Additionally, it is important to estimate how important or central a sentence is within the document. These two components, the context and the centrality of the sentences, are studied in this thesis following a formal framework based on Statistical Language Models. In this respect, we demonstrate that these components yield to improvements in current sentence retrieval methods.

In this thesis we work with collections of sentences that were extracted from news. News not only explain facts but also express opinions that people have about a particular event or topic. Therefore, the proper estimation of which passages are opinionated may help to further improve the estimation of relevance for sentences. We apply a formal methodology that helps us to incorporate opinions into standard sentence retrieval methods. Additionally, we propose simple empirical alternatives to incorporate query-independent features into sentence retrieval models. We demonstrate that the incorporation of opinions to estimate relevance is an important factor that makes sentence retrieval methods more effective. In the course of our study, we also analyze query-independent features based on sentence length and named entities.

The combination of the context-based approach with the incorporation of opinion-based features is straightforward. We study how to combine these two approaches and the impact of such combination. We demonstrate that context-based models are implicitly promoting sentences with opinions and, therefore, opinion-based features do not help to further improve context-based methods.

The second part of this thesis is dedicated to novelty detection at sentence level. Because novelty is actually dependent on a retrieval ranking, we consider here two approaches: a) the perfect-relevance approach, which consists of using a ranking where all sentences are relevant (this is an ideal approach); and b) the non-perfect relevance approach, which consists of applying first a sentence retrieval method (therefore, the ranking may contain sentences that are not relevant).

We first study which baseline performs the best and, next, we propose a number of variations. One of the mechanisms proposed is based on vocabulary pruning. We demonstrate that considering terms from the top ranked sentences in the original ranking helps to guide the estimation of novelty. The application of Language Models to support novelty detection is another challenge that we face in this thesis. We apply different smoothing methods (Dirichlet and Jelinek-Mercer) in the context of alternative mechanisms to detect novelty (Aggregate and Non-Aggregate Models). Additionally, we test a mechanism based on mixture models that uses the Expectation-Maximization algorithm to obtain automatically the novelty score of a sentence.

In the last part of this work we demonstrate that most novelty methods lead to a strong re-ordering of the initial ranking. However, we show that the top ranked sentences in the initial list are usually novel and re-ordering them is often harmful. Therefore, we propose different mechanisms that determine the position threshold where novelty detection should be initiated. In this respect, we consider query-independent (a fixed position for all queries) and query-dependent approaches (cluster-based and normalized-score approaches).

Summing up, we identify important limitations of current sentence retrieval and novelty methods and, along this thesis, we propose alternative methods that are novel and effective.