Evaluation Methods for Wikification

Denis Fedorenko, Andrey Sysoev, Nikita Astrakhantsev

Institute for System Programming of the Russian Academy of Sciences
{fedorenko,sysoev,astrakhantsev}@ispras.ru

Abstract. This paper presents a more strict evaluation technique for the wikification task. The main difference from the previous evaluation methods is that it also takes into account the positions of the extracted terms and considers the terms having no appropriate concepts among candidates or in whole Wikipedia. Our method shows lower values of precision and recall, but represents the quality of wikification results more actually.

1 Introduction

Wikification is an automatic process of identifying meaningful phrases, or *terms*, in a given text and disambiguating them to the corresponding Wikipedia articles, or *concepts*. Wikification is one of the most important components of different NLP tasks where the conceptual representation of texts is needed.

Most of the evaluation methods for wikification do not fully represent the actual quality of the end results. For example, in studies [1], [2] the authors compare a set of Wikipedia titles determined by the wikification algorithm with the gold standard, ignoring the terms for which these titles are assigned. In [3] the same technique is used, but the set of output titles is additionally constrained by the terms appeared in the gold standard. Moreover, these studies ignore the terms having no appropriate concepts among candidate titles or in whole Wikipedia.

In the same time, for such NLP tasks as automatic text markup and ontology enrichment both terms and concepts determined in the document are important and hence should be evaluated. It motivates us to propose a more strict evaluation technique for the wikification task. The description of this technique and the obtained results are presented in the next section.

2 Evaluating Wikification

Let $DP_{expected}$, DP_{actual} are sets of disambiguated phrases of the gold standard and extracted by the algorithm, respectively. *Disambiguated phrase* is a pair *(term, concept)*, where the term is the position of the meaningful phrase in the text and the concept is either identifier of some Wikipedia article or special NOT_IN_KB value denoting that the term has no appropriate meaning in whole

Wikipedia. Then, the overall quality of the algorithm is estimated by the classical precision and recall metrics:

$$P = \frac{|DP_{expected} \cap DP_{actual}|}{|DP_{actual}|} \quad (1)$$

$$R = \frac{|DP_{expected} \cap DP_{actual}|}{|DP_{expected}|} \quad (2)$$

To test the approach, we implemented the wikification system "GLOW" [3] and compared results with the evaluation methods employed in the previous papers. We used two datasets for the evaluation: the first one, MODIS, is a collection of 131 texts of general domain, and the second one, Board Games, contains 35 domain-specific texts. In these datasets all possible non-overlapping terms occurred in the texts are manually annotated, including those that have no appropriate meanings among candidates or in whole Wikipedia.

The results are presented in Tables 1, 2. As we can see, there is a big difference between the values of quality metrics, especially on domain-specific texts; our method shows the lowest results. It leads to the conclusion that taking into account the positions of terms and NOT_IN_KB concepts significantly decrease the values of metrics, but represent the quality of wikification more actually.

	Ratinov			**Milne**			**Our approach**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
MODIS	74%	74%	74%	60%	46%	52%	56%	37%	44%
Board Games	47%	69%	56%	45%	41%	49%	41%	19%	26%

Table 1. Results for wikification. Ratinov, Milne are the evaluation methods used in [3], [1], respectively

	Ratinov,Milne			**Our approach**		
	P	**R**	**F1**	**P**	**R**	**F1**
MODIS	82%	82%	82%	76%	69%	72%
Board Games	75%	75%	75%	49%	40%	44%

Table 2. Results for disambiguation only (the terms are passed to the algorithm as input)

3 Conclusions and Future Work

In this study we proposed a more strict evaluation technique for the wikification task. The main difference from the previous methods is that it takes into account the positions of terms and considers the terms having no appropriate concepts among candidate titles or in whole Wikipedia. Nevertheless, more comprehensive study is needed to detect the most actual errors of existing wikification methods.

References

1. Milne D., Witten I. H. Learning to link with wikipedia.
2. Ferragina, P. et al. Tagme: on-the-fly annotation of short text fragments.
3. Ratinov L. et al. Local and global algorithms for disambiguation to wikipedia.