Classification-based scientific term detection in patient information

Veronique Hoste, Klaar Vanopstal, Els Lefever, Isabelle Delaere

LT³, University College Ghent

Despite the legislative efforts to improve the readability of patient information, different surveys have shown that respondents still feel distressed by reading the patient information leaflet. One of the main sources of distress is the use of scientific terminology. In order to assess the scale of the problem, we collected a Dutch-English parallel corpus of European Public Assessment Reports (EPARs) which was annotated by two linguists. This corpus was used to evaluate and train an automatic approach to scientific term detection.

As an alternative to the dictionary-based approaches, which suffer from low coverage, we present a classification-based approach relying on a wide variety of information sources, such as local context and lexical information, termhood and unithood information, cognate identification and morphological information. We present the experimental findings and show some first results of the automatic replacement of the detected terms by their popular counterpart.