# Assignment week 7 Text mining - Tanja Crijns s4204999

**1. What information from the Yahoo data do the authors use as labels for the correctness ofanswers?**

The authors state:

> The candidate answer set for a given question is composed by one positive example, i.e., its corresponding best answer, and as negative examples all the other answers retrieved in the top N by the retrieval component

**2. Explain how the authors address the problem of the vocabulary gap between question and answer without using external knowledge bases such as WordNet.**

The authors mention a possible solution:

> One way to address this problem is to learn question-to-answer transformations using a translation model (Berger et al., 2000; Echihabi and Marcu, 2003; Soricut and Brill, 2006; Riezler et al., 2007).

Their own approach, based on this solution, is as follows:

> In our model, we incorporate this approach by adding the probability that the question Q is a translation of the answer A, P(Q|A), as a feature.

Instead of external knowledge, they use translation features. This means that they use a maximum likelihood estimation in a Bayesian framework that the Question Q is a Translation of the answer A.

**3. Give a definition in your own words of what the authors call 'Recall@N'. Do you think that the term 'Recall@N' is a good descriptor of this definition? Please motivate your answer.**

The term 'Recall@N' refers to the fraction of sets of top N results that contain the correct answer. This does not comply with the ordinary definition of recall, which has the following definition: Recall is the number of relevant retrieved documents divided by the total number of relevant documents. I do not think recall is the best descriptor of the definition.

**4. To what extent would the feature set in this paper be applicable to factoid questions? Which of the features would not be?**

The feature set consists of similarity features, translation features, density/frequency features and web correlation features. An answer to a factoid questions is usually short, like for example: 'Q: When was Tanja Crijns born? A: 1994'.
It seems similarity features would be applicable to both short and long entries, as far as my knowledge about them stretches. This goes the same for translation features and web correlation features. I do not see why they would not work for factoid questions. Density/frequency features however work better when the answer is longer.