ANLP Seminar 3: Language Modelling

Julie Weeds

February 11, 2019

Zweig and Burges (2011) introduce an evaluation task known as the Microsoft Research Sentence Completion Challenge. We will be looking at this task and its potential to evaluate language models. There are 1040 *questions* in the dataset, an example of which is given below:

> **Was she his [client ‖ musings ‖ discomfiture ‖ choice ‖ opportunity], his friend or his mistress?**

1. Explain what the task is, for a human or a computer system, for a question, as presented above. In the above example, what knowledge is needed in order to choose the correct answer?
2. How was the dataset created? How and why were the incorrect answers selected in the way they were?
3. How is performance measured on this task? What score must a method achieve to be better than the *baseline* of random guessing?
4. What are the advantages and disadvantages of this evaluation task compare to ones based on human synonymy judgements such as WordSim353?
5. How does the simple 4-gram model work?
6. What do you understand by a *smoothed* n-gram model?
7. Explain the method based on latent semantic analysis similarity. Why do you think this does better than the n-gram methods?
8. How do you think you could do better on this task (without asking humans to help!)?

References

Geoffrey Zweig and Christopher Burges. 2011. The microsoft research sentence completion challenge. Technical report, Microsoft Research, December.