Leveraging Lexical Substitutes for Unsupervised Word Sense Induction

Word sense induction is the most prominent unsupervised approach to lexical disambiguation. It clusters word instances, typically represented by their bag-of-words contexts. Therefore, uninformative and ambiguous contexts present a major challenge. In this paper, we investigate the use of an alternative instance representation based on lexical substitutes, i.e., contextually suitable, meaning-preserving replacements. Using lexical substitutes predicted by a state-of-the-art automatic system and a simple clustering algorithm, we outperform bag-of-words instance representations and compete with much more complex structured probabilistic models. Furthermore, we show that an oracle based on manually-labeled lexical substitutes yields yet substantially higher performance. Taken together, this provides evidence for a complementarity between word sense induction and lexical substitution that has not been given much consideration before.