Segmentation of existing methods for text segmentation

segmentation as to divide a text into a sequenceof terms. Statistical approaches, such as N-gram Model 21, 22, 23, calculate the frequencies of words co-occurring as neighbors in a trainingcorpus. When the frequency exceeds a prede? ned threshold, the correspondingneighboring words can be treated as a term. Vocabulary-based approaches 18, 19, 20 extract terms by checking for their existence or frequency in a prede? nedvocabulary. The most obvious drawback of existing methods for text segmentationis that they only consider surface features and ignore the requirement ofsemantic coherence within a segmentation. This might lead to incorrectsegmentations as described in Challenge 1. To this end, we propose to exploitcontext semantics when conducting text segmentation. POS tagging.

POS taggingdetermines lexical types (i. e., POS tags) of words in a text. Rule-based POStaggers attempt to assign POS tags to unknown or ambiguous words based on alarge number of hand-crafted 10, 11 or automatically learned 12, 13 linguisticrules. Statistical POS taggers avoid the cost of constructing tagging rules bybuilding a statistical model automatically from a corpora and labeling untaggedtexts based on those learned statistical information. Mainstream statisticalPOS taggers employ the well-known Markov Model 14, 15, 16, 17 whichlearns both lexical probabilities and sequential probabilities from a labeledcorpora and tags a new sentence by searching for tag sequence that maximizesthe combination of lexical and sequential probabilities. Note that bothrule-based and statistical POS taggers rely on the assumption that texts arecorrectly structured which, however, is not always the case for short texts. More importantly, existing methods only considers lexical features and ignoresword semantics.

This might lead to mistakes, as illustrated in Challenge 3. Ourwork attempts to build a tagger which considers both lexical features andunderlying semantics for type detection. Semantic labeling. Semantic labelingdiscovers hidden semantics from a natural language text. Named entityrecognition (NER) locates named entities in a text and classi? es them intoprede? ned categories (e.

g., persons, organizations, locations, etc.) usinglinguistic grammar-based techniques as well as statistical models like CRF 1and HMM 2. Topic models 3 attempt to recognize “ latent topics”, which arerepresented as probabilistic distributions on words, based on observablestatistical relations between texts and words.

Entity linking 5, 6, 7, 8 employs existing knowledgebases and focuses on retrieving “ explicit topics” expressed as probabilistic distributions on the entire knowledge base. Despitethe high accuracy achieved by existing work on semantic labeling, there arestill some limitations. First, categories, “ latent topics”, and “ explicittopics” are different from human-understandable concepts.

Second, short textsdo not always observe the syntax of a written language which, however, is anindispensable feature for mainstream NER tools. Third, short texts do notcontain suf? cient content to support statistical models like topic models. Thework most related to ours are conducted by Song et al. 19 and Kim et al. 20respectively, which also represent semantics as concepts. 19 employs theBayesian Inference mechanism to conceptualize instances and short texts, andeliminates instance ambiguity based on homogeneous instances.

Kim et al. 20captures semantic relatedness between instances using a probabilistic topicmodel (i. e., LDA), and disambiguates instances based on related instances.

Inthis work, we observe that other terms, such as verbs, adjectives, andattributes, can also help with instance disambiguation. We incorporate typediscernment in to our framework for short text understanding of conductinstance disambiguate based on various types of context information.