DISTINGUISHING COMMON AND PROPER NOUNS IN MACHINE TRANSLATION

Sharipova Aziza Abdumanapovna

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi head of the Department of Foreign Languages PhD, dotsent.

Matyakubova Noila Shakirjonovna

Uzbek State World Languages University

Abstract: This article describes a number of techniques for automatically deriving lists of common and proper nouns, and shows that the distinction between two types of nouns can be made automatically by using a vector space model learning algorithm.

Key words: Proper names, taggers, Ambiguous nouns, Automatic speech recognition, a vector space model technique.

Introduction.

Proper names constitute a difficult problem in machine translation (MT). In Uzbek and English languages, the default is that if a word is written with a capital initial letter, it is a proper name. It can be either at the beginning of the sentence or any other part of the sentence, which still complicates the matter. Moreover, some nouns are homographs (they have the same written form, but different meaning) which can be used to denote either a common or proper noun, for instance the word apple in the following examples: Apple designs and creates iPod or An apple is a fruit of the apple tree . The common and proper uses are not always as clearly distinct as in this example; for example, a specific instance of a common noun, e.g., District Court turns court into a proper noun. While heuristically, proper nouns often start with a capital letter in English, capitalization can be inconsistent, incorrect or omitted, and the presence or absence of an article cannot be relied on.

The problem of distinguishing between common and proper usages of nouns has not received much attention within language processing, despite being an important component for many tasks including machine translation (Lopez, 2008; Hermjakob et al., 2008), sentiment analysis (Pang and Lee, 2008; Wilson et al., 2009) and topic tracking (Petrovic et al., ´ 2010). Approaches to the problem also have applications to tasks such as web search (Chen et al., 1998; Baeza-Yates and Ribeiro-Neto, 2011), and case restoration such as automatic speech recognition output, but frequently involve the manual creation of a list of proper nouns, which suffer not only from omissions but also often do not allow the listed words to assume their common role in text. It can

ISSN 2181-2632 www.involta.uz

be deduced that it is difficult to automatically construct lists of ambiguous nouns but also that they can be distinguished effectively using standard features from Word Sense Disambiguation.

Approaches to resolve ambiguity

A solution which comes first into mind is to list all proper names into the analysis system, either to the morphological analyser or to a post-processing module. In practice, however, such a list would be always defective, because new names appear continually. A better approach is: list only such names that need translation or such semantic labelling that cannot be determined on the basis of the word form itself.

Also the need for adding semantic information may give reason for listing. For example, such distinctions as animate/non-animate, male/female and singular /plural may be reason for listing proper names. For the rest of non-sentence-initial but capital-initial words can be given the default interpretation of proper name. They do not need listing anywhere. If such a word has also an ordinary meaning, the analysis system produces at least two interpretations, one for a proper noun and one or more for ordinary words (note that the word can be a noun or a form of some other word class).

When considering disambiguation, it would be tempting to select the proper name interpretation for words that are capital-initial but not sentence-initial. However, the situation is not that simple. Texts often have capital-initial words inside the sentence, for stylistic and whatever reasons. They should be kept separate from true proper names. The solution is to mark with a special tag such words that are also potential proper names. Such tags can be added to the morphological lexicon or produce in a post-editing process. Using this method we can control whether the word is a potential proper name. The words that do not have that tag will be interpreted as normal words although they would be capital-initial and non-sentence-initial. This requires keeping control of the list of words that may have both interpretations. The problem that remains is how to handle ambiguous capital-initial words that are at the same time sentence-initial. In this position all words are capital-initial. It is hardly possible to solve this problem exhaustively. Probability measures would, however, bring satisfactory results.

Generating Lists of Nouns

To our knowledge, no comprehensive list of common nouns with proper noun usage is available. We develop a number of heuristics to generate such lists automatically.

Part of speech tags. A number of part of speech taggers assign different tags to common and proper nouns. Ambiguous nouns are identified by tagging a corpus and extracting those that have had both tags assigned, together with the frequency of occurrence of the common/proper usage. The CLAWS (Garside, 1987) and the RASP taggers were applied to the British National Corpus to generate the lists British National Corpus claws and British National Corpus rasp respectively. In addition the RASP tagger was also run over the 1.75 billion word Gigaword corpus (Graff, 2003) to extract the list Gigaword.

Wikipedia includes disambiguation pages for ambiguous words which provide information about their potential usage. Wikipedia pages for nouns with senses (according to the disambiguation page) in a set of predefined categories were identified to form the list Wikipedia.

ISSN 2181-2632 www.involta.uz

Named entity recognition The Stanford Named Entity Recogniser was run over the BNC and any nouns that occur in the corpus with both named entity and non-named entity tags are extracted to form the list Stanford.

We cast the problem of distinguishing between common and proper usages of nouns as a classification task and develop the following approaches. 3.1 Most frequent usage A naive baseline is supplied by assigning each word its most frequent usage form (common or proper noun). The most frequent usage is derived from the training portion of labeled data.

Vector Space Model (VSM).

Distinguishing between common and proper nouns can be viewed as a classification problem. Treating the problem in this manner is reminiscent of techniques commonly employed in Word Sense Disambiguation (WSD). Our supervised approach is based on an existing WSD system (Agirre and Martinez, 2004) that uses a wide range of features:

• Word form, lemma or Parts of Speech bigrams and trigrams containing the target word.

• Preceding or following lemma (or word form) content word appearing in the same sentence as the target word.

• High-likelihood, salient, bigrams.

• Lemmas of all content words in the same sentence as the target word.

• Lemmas of all content words within a ±4 word window of the target word.

• Nonstop word lemmas which appear more than twice throughout the corpus.

Each occurrence of a common / proper noun is represented as a binary vector in which each position indicates the presence or absence of a feature. A centroid vector is created during the training phase for the common noun and the proper noun instances of a word. During the test phase, the centroids are compared to the vector of each test instance using the cosine metric, and the word is assigned the type of the closest centroid.

The vector space model outperforms other approaches on both corpora. Performance is particularly high when capitalisation is included .However, this approach still outperforms the baseline without case information, demonstrating that using this simple approach is less effective than making use of local context.

Automatic speech recognition.

Most automatic speech recognition (ASR) systems do not provide capitalization. However, our system does not rely on capitalization information, and therefore can identify proper / common nouns even if capitalization is absent. Also, once proper nouns are identified, the system can be used to restore case – a feature which allows an evaluation to take place on this dataset. We use the TDT2 Test and Speech corpus (Cieri et al., 1999), which contains ASR and a manually transcribed version of news texts from six different sources, to demonstrate the usefulness of this system for this task. The ASR corpus is restricted to those segments which contain an equal number of target word occurrences in the ASR text and the manually

ISSN 2181-2632 www.involta.uz

transcribed version, and all such segments are extracted. The gold standard, and the most frequent usage, are drawn from the manually transcribed data.

Conclusion.

We have described in detail how to generate lists of common and proper nouns automatically using a number of different techniques. A vector space model technique for distinguishing common and proper nouns is found to achieve high performance when evaluated on the BNC. This greatly outperforms, due to its better adaptability to sparse training data. Automatic speech recognition systems based evaluations also demonstrate the system’s performance and as a side effect the system could serve as a technique for automatic case restoration.

References.

1. Baeza-Yates, R. and Ribeiro-Neto, B. (2011). Modern Information Retrieval: The Concepts and Technology Behind Search. Addison Wesley Longman Limited, Essex.
2. Bird, S., Klein, E., and Loper, E. (2009). Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. O’Reilly.
3. Brants, T. and Franz, A. (2006). Web 1T 5-gram v1. Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the RASP system. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions.
4. Chen, H., Huang, S., Ding, Y., and Tsai, S. (1998). Proper name translation in cross-language information retrieval. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1, pages 232–236, Montreal, Canada.
5. Cieri, C., Graff, D., Liberman, M., Martey, N., and Strassel, S. (1999). The TDT-2 text and speech corpus. In Proceedings of DARPA Broadcast News Workshop, pages 57–60.
6. Fellbaum, C., editor (1998). WordNet: An Electronic Lexical Database and some of its Applications. MIT Press, Cambridge, MA.
7. Lopez, A. (2008). Statistical machine translation. ACM Computing Surveys, 40(3):1–49.
8. Garside, R. (1987). The CLAWS word-tagging system. In Garside, R., Leech, G., and Sampson, G., editors, The Computational Analysis of English: A Corpusbased Approach. London: Longman.
9. Graff, D. (2003). English Gigaword. Technical report, Linguistic Data Consortium.

ISSN 2181-2632 www.involta.uz