Target Word Selection Using WordNet and Data-Driven Models in Machine Translation

Yuseop Kim[1], Jeong-Ho Chang[2], and Byoung-Tak Zhang[2]

[1] Ewha Institute of Science and Technology, Ewha Woman's University, Seoul, Korea 120-750
{yskim01}@ewha.ac.kr***
[2] School of Computer Science and Engineering, Seoul National University, Seoul, Korea 151-744
{jhchang, btzhang}@bi.snu.ac.kr †

Collocation information plays an important role in target word selection of machine translation. However, a collocation dictionary fulfills only a limited portion of selection operation because of data sparseness. To resolve the sparseness problem, we proposed a new methodology that selects target words after determining an appropriate collocation class by using a inter-word semantic similarity. We estimate the similarity by computing semantic distance of two synsets in WordNet and term-to-term similarity in data-driven models. In WordNet, semantic similarity between two word can be calculated by adapting a reciprocal of the Semantic Distance (SD). For the calculation of the SD, each synset in WordNet is assigned an M-value. The M-value is computed as follows: $M\text{-value} = \frac{radix}{sf^p}$, where $radix$ is an initial M-value, sf is a scale factor, and p is the number of edges from the root to the synset. As the data-driven models, we utilize Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis(PLSA), a probabilistic application of LSA. LSA applies singular value decomposition (SVD) to the matrix. SVD is a form of factor analysis and is defined as $A = U\Sigma V^T$,where Σ is a diagonal matrix composed of nonzero eigen values of AA^T or A^TA, and U and V are the orthogonal eigenvectors associated with the r nonzero eigenvalues of AA^T and A^TA, respectively. The term-to-term similarity is based on the inner products between two row vectors of A, $AA^T = U\Sigma^2U^T$. And To compute the similarity of w_1 and w_2 in PLSA, $P(z|w_1)P(z|w_2)$ should be approximately computed with being derived from $P(z|w) = \frac{P(z)P(w|z)}{\sum_z P(z)P(w|z)}$, where z represents contexts. For experiments, we implemented three similarity measurements applying to WordNet, LSA and PLSA and we used TREC data(AP news in 1988). We could obtain up to 18% accuracy improvement from suggested approaches, comparing to direct matching to a collocation dictionary.

*** He is supported by Brain Korea 21 project performed by Ewha Institute of Science and Technology

† They are supported by Brain Tech. project from Korean Ministry of Science and Technology

M. Ishizuka and A. Sattar (Eds.): PRICAI 2002, LNAI 2417, p. 607, 2002.