中央研究院 資訊科學研究所
Institute of Information Science, Academia Sinica

Distinguished Lecture Series

Towards Google-like Search on Spoken Documents with Zero Resources

How to get something from nothing in a language that you've never heard of

Monday, September 17th, 2012 10:00am
Auditorium 106 at New IIS Building

Kenneth Church

Researcher, IBM

Abstract

There is considerable interest in interdisciplinary combinations of automatic speech recognition (ASR), machine learning, natural language processing, text classification and information retrieval (IR). Many of these systems, especially ASR, are often based on large (expensive) linguistic resources. When we have resources, we should use them. But when we don't, we can still do much of what you can do with bags of words (BOWs), even when many/most/all terms are out-of-vocabulary (OOV).

The proposed method finds long (~ 1 sec) repetitions in speech, and clusters them into pseudo-terms (roughly phrases such as "Johns Hopkins University"). Documents are represented as bags of pseudo-terms (BOPs) instead of BOWs. Standard IR features such as term frequency (tf) and document frequency (df) can be computed over BOPs instead of BOWs. IR tasks such as document clustering and classification can be reformulated to use BOPs instead of BOWs.

For more information: http://www.iis.sinica.edu.tw/