Linguists use the Bible to develop language technology for small languages

8 September 2015

If you speak English or another big language, you can talk to your mobile phone, use search engines, and get machine translation systems to do your translations for you. This has been made possible because English is a huge language with a great number of resources that linguists employ to develop language technology. People who speak Faroese, Welsh or Galician are less fortunate.

"When we develop machine translation systems and search engines, we usually feed huge amounts of manually annotated texts that contain information about the function and meaning of individual words into a computer. For historical reasons, these texts have primarily been newspaper articles in English and other big languages. We do not have access to similarly annotated texts in smaller languages like Faroese, Welsh, Galician and Irish, or even a major African language like Yoruba which is spoken by 28 million people," says Professor Anders Søgaard from the University of Copenhagen.

Anders Søgaard and his colleagues from the project LOWLANDS: Parsing Low-Resource Languages and Domains are utilising the texts which were annotated for big languages to develop language technology for smaller languages, the key to which is to find translated texts so that the researchers can transfer knowledge of one language's grammar onto another language:

"The Bible has been translated into more than 1,500 languages, even the smallest and most 'exotic' ones, and the translations are extremely conservative; the verses have a completely uniform structure across the many different languages which means that we can make suitable computer models of even very small languages where we only have a couple of hundred pages of biblical text," Anders Søgaard says and elaborates:

"We teach the machines to register what is translated with what in the different translations of biblical texts, which makes it possible to find so many similarities between the annotated and unannotated texts that we can produce exact computer models of 100 different languages - languages such as Swahili, Wolof and Xhosa that are spoken in Nigeria. And we have made these models available for other developers and researchers. This means that we will be able to develop language technology resources for these languages similar to those which speakers of languages such as English and French have."

Anders Søgaard and his colleagues have recently presented their results in the article "'If you all you have is a bit of the Bible' at the prestigious conference Annual Meeting of the Association of Computational Linguistics.

Wikipedia as universal dictionary

The user-driven online encyclopaedia Wikipedia has also proved to be a highly useful source for the researchers who use its texts to develop language resources for languages where people do not have access to the new language technologies. Wikipedia contains over 35 million articles, but it is

the fact that as many as 129 languages are represented by more than 10,000 articles each that the researchers find interesting as many articles concern the same concepts and topics.

"This allows us to do what we call 'inverted indexing' which means that we use the concept that the Wikipedia articles is about to describe the words used in the articles on the concept in different languages. We usually use the words to describe the concept but here we do it in reverse order," Anders Søgaard explains and continues:

"If the English word 'glasses' appears in the English Wikipedia entry on Harry Potter, and the German word 'Brille' is used in the equivalent German entry, it is very likely that the two words will be represented in a similar fashion in our models which form the basis of e.g. machine translation systems. And the advantage of this model is that it can be applied to 100 different languages at the same time, including many languages that have previously been denied the language technology resources that we use every day."

The method is described in the article 'Inverted indexing for cross-lingual NLP' which Anders Søgaard wrote together with researchers from Google London. The article was also presented at the Annual Meeting of the Association of Computational Linguistics.

More information: Annual Meeting of the Association of Computational Linguistics: aclweb.org/anthology/P15-2044

Provided by University of Copenhagen