THESIS PROPOSAL

Identifying Different Translations of the Same Book

Background

When Storytel receives e-books from publishers, they receive both the book content and metadata. The metadata files follow a standard called ONIX. There is very limited support in this standard to represent the fact that two books are actually the same literary work but just different translations.

Project description

The purpose of this thesis project is to explore unsupervised methods for identifying publications that represent different translations of the same book. The main research question is whether this problem can be attacked using polylingual topic models of the books' textual contents, as has been proposed by Krstovski and Smith (2013). The basic approach could be extended to capture more book-specific text features (such as character names). The final approach will be evaluated by staff at Storytel.

An initial problem when using a polylingual topic model is that there need to be translated books to anchor topical structure between languages. To address this problem the project will study the possibility to use either some separate parallel corpus (such as Wikipedia) or some machine translation system (such as Google Translate) for topic anchoring.

Kriste Krstovski and David A. Smith. Online Polylingual Topic Models for Fast Document Translation Detection. In WMT@ ACL, pages 252–261, 2013.

Customer

Storytel, Stockholm

Contact

Marco Kuhlmann, marco.kuhlmann@liu.se

Student profile

Background knowledge in statistical modelling of natural language and text mining. Interest in language and translation.