JTDH 2016, Ljubljana, Slovenia

University of Novi Sad, Faculty of Technical Sciences
AlfaNum Speech Technologies

The Use of Semantic Word Classes in Document Classification

Stevan Ostrogonac / Branislav Popović / Milan Sečujski

TR 32035 / E! 9944

Introduction

- **Document classification** and **topic modelling** represent some of the biggest challenges in natural language processing and information retrieval.
- Many of the techniques developed for these purposes are **language-independent**.
- Language resources are needed for each language, along with domain-specific data sets for particular applications.
  - Every new language introduces a specific set of problems.
- The problem of **data sparsity** in document classification is addressed for under-resourced, highly inflective languages.

Introduction

- Serbian language is considered, but the method is applicable to other languages as well.
- The approach includes
  - training a **language model** (LM) on a large textual corpus
  - using the LM to create **semantic word classes**
  - using the extracted semantic information to obtain more robust **document classifier**.
- Latent Dirichlet Allocation (**LDA**) can be used as a topic model, as well as its variants or other types of topic models.

Semantic Information Extraction

- The language model was trained using a textual corpus for Serbian that contains
  - over 20 million word tokens
  - ~ 360 thousand word types
  - ~ 180 thousand lemmas
  - ~ 1000 morphologic classes
- The LM was **lemma-based**.
  - Morphologic information is available for the Serbian language (Sečujski, 2002) and could be restored after semantic lemma classes were derived.

Semantic Information Extraction

- The semantic classes were created by applying a **greedy clustering algorithm** (Mikolov, 2012) to the **lemmatized** textual corpus.
- The clustering algorithm leans on the probabilities obtained from the LM for hypotheses created by replacing a lemma with other lemmas from the dictionary.
- The lemmas for which the replacement causes the smallest change in probabilities are likely to be semantically similar to the original word.
- After the entire corpus is processed, and morphologic information is restored to derive words from lemmas, **semantic word classes** are created.

Semantic Information Extraction

- The parameters for clustering should be **fine-tuned** iteratively by observing the results and adjusting the values.
  - The classes are optimized for a particular application.
- Each **semantic class** can represent
  - only synonyms
  - all the words that can be placed in certain positions within sentences and result in semantically correct sentences
  - or something in between.

Semantic Word Classes in LDA

- **LDA** is a generative model which can be used for **document classification**.
- In LDA, a document is considered to be a mixture of a number of **topics**, which is similar to the bag of words concept.
- Each word may belong to many topics, to each with a certain probability.
- In order to define those probabilities and the topics themselves, a great amount of data is needed.

Semantic Word Classes in LDA

- One of the most popular document classification tasks is **e-mail classification** into regular messages and spam.
- Two spam messages can contain similar or the same topics, but consist of very different sets of words.
  - *“Buy now at lower price and enjoy the trip!”*
  - *“Purchase immediately, experience an exciting travel with our discount!”*
- This problem is emphasized in **highly inflective** languages.
- Even though textual data of specific content may not be enough to train highly accurate classifiers, other textual resources can be used to obtain additional information.

Semantic Word Classes in LDA

- Semantic classes derived from a large textual corpus which contains many **different types of documents** can be used to make a document classifier **more robust**.
- By using semantic class IDs instead of words, an LDA can model topics quite well even with a small amount of **application-specific data**.
  - For each word that is observed within the training data set, an entire semantic class is included in the modelling process.
  - Semantic classes may be grouped manually, or by applying a rule-based approach (e.g., word-stem derivation) in the case that morphologic dictionary is not available.

Semantic Word Classes in LDA

- **Semantic word clustering** insures that words with the same meaning but different morphological features are grouped together.
  - Therefore, eliminates morphology as a cause of **data sparsity** in topic modelling.
- Semantic classes include words with similar meaning, which **reduces the number of topics** to be modelled, resulting in **more accurate** topic representations.
- The application of the described approach is far more broad and includes different information retrieval tasks.

Further Research and Application

- Semantic word clustering can be improved by implementation of a probabilistic approach (i.e., words that belong to more than one semantic class with corresponding probabilities).
- Two semantic classes
  - **A** = {malaria, flu, meningitis, AIDS,…}
  - **B** = {drug, medicine, therapy, cure,…}
- could be highly semantically correlated, but this information is not extracted.
- Obtaining **higher-level semantic information** requires wider context analysis, which will be the main topic of further research.

Further Research and Application

- Applications of the extracted semantic information are numerous. This research represent the basis for creation of **advanced dialogue systems**, able to mimic natural dialogue.
- The most important pursuit in this area would be to develop the possibility of determining the **meaning** of a word that a dialogue system has not seen before.

University of Novi Sad, Faculty of Technical Sciences
AlfaNum Speech Technologies

The Use of Semantic Word Classes in Document Classification

Stevan Ostrogonac / Branislav Popović / Milan Sečujski

Thank you!