1: Predicting the Quality of Questions on Stackoverflow

Community Question Answering websites (CQA) have a growing popularity as a way of providing and searching of information. CQA attract users as they provide a direct and rapid way to find the desired information. As recognizing good questions can improve the CQA services and the user's experience, the current study focuses on question quality instead. Specifically, we predict question quality and investigate the features which influence it. The influence of the question tags, length of the question title and body, presence of a code snippet, the user reputation and terms used to formulate the question are tested. For each set of dependent variables, Ridge regression models are estimated. The results indicate that the inclusion of terms in the models improves their predictive power. Additionally, we investigate which lexical terms determine high and low quality questions. The terms with the highest and lowest coefficients are semantically analyzed. The analysis shows that terms predicting high quality are terms expressing, among others, excitement, negative experience or terms regarding exceptions. Terms predicting low quality questions are terms containing spelling errors or indicating off-topic questions and interjections

2: Knowledge Graph Embedding via Dynamic Mapping Matrix

Knowledge graphs are useful resources for numerous AI applications, but they are far from completeness. Previous work such as TransE, TransH and TransR/CTransR regard a relation as translation from head entity to tail entity and the CTransR achieves state-of-the-art performance. In this paper, we propose a more fine-grained model named TransD, which is an improvement of TransR/CTransR. In TransD, we use two vectors to represent a named symbol object (entity and relation). The first one represents the meaning of a(n) entity (relation), the other one is used to construct mapping matrix dynamically. Compared with TransR/CTransR, TransD not only considers the diversity of relations, but also entities. TransD has less parameters and has no matrix-vector multiplication operations, which makes it can be applied on large scale graphs. In Experiments, we evaluate our model on two typical tasks including triplets classification and link prediction. Evaluation results show that our approach outperforms state-of-the-art methods

3: Improving Web 2.0 Opinion Mining Systems Using Text Normalisation Techniques

A basic task in opinion mining deals with determining the overall polarity orientation of a document about some topic. This has several applications such as detecting consumer opinions in online product reviews or increasing the effectiveness of social media marketing campaigns. However, the informal features of Web 2.0 texts can affect the performance of automated opinion mining tools. These are usually short and noisy texts with presence of slang, emoticons and lexical variants which make more difficult to extract contextual and semantic information. In this paper we demonstrate that the use of lexical normalisation techniques can be used to enhance polarity detection results by replacing informal lexical variants with their canonical version. We have carried out several polarity classification experiments using English texts from different Web 2.0 genres and we have obtained the best result with microblogs where normalisation contribution to the classification model can be up to 6.4

4: PanLex: Building a Resource for Panlingual Lexical Translation

PanLex, a project of The Long Now Foundation, aims to enable the translation of lexemes among all human languages in the world. By focusing on lexemic translations, rather than grammatical or corpus data, it achieves broader lexical and language coverage than related projects. The PanLex database currently documents 20 million lexemes in about 9,000 language varieties, with 1.1 billion pairwise translations. The project primarily engages in content procurement, while encouraging outside use of its data for research and development. Its data acquisition strategy emphasizes broad, high-quality lexical and language coverage. The project plans to add data derived from 4,000 new sources to the database by the end of 2016. The dataset is publicly accessible via an HTTP API and monthly snapshots in CSV, JSON, and XML formats. Several online applications have been developed that query PanLex

data. More broadly, the project aims to make a contribution to the preservation of global linguistic diversity.

5: Sockpuppet Detection in Wikipedia: A Corpus of Real-World Deceptive Writing for Linking Identities

This paper describes a corpus of sockpuppet cases from Wikipedia. A sockpuppet is an online user account created with a fake identity for the purpose of covering abusive behavior and/or subverting the editing regulation process. We used a semi-automated method for crawling and curating a dataset of real sockpuppet investigation cases. To the best of our knowledge, this is the first corpus available on real-world deceptive writing. We describe the process for crawling the data and some preliminary results that can be used as baseline for benchmarking research. The dataset has been released under a Creative Commons license from our project website (http://docsig.cis.uab.edu/tools-and-datasets/).

6: Question Answering over Freebase with Multi-Column Convolutional Neural Networks

Answering natural language questions over a knowledge base is an important and challenging task. Most of existing systems typically rely on hand-crafted features and rules to conduct question understanding and/or answer ranking. In this paper, we introduce multi-column convolutional neural networks (MCCNNs) to understand questions from three different aspects (namely, answer path, answer context, and answer type) and learn their distributed representations. Meanwhile, we jointly learn low-dimensional embeddings of entities and relations in the knowledge base. Question-answer pairs are used to train the model to rank candidate answers. We also leverage question paraphrases to train the column networks in a multi-task learning manner. We use FREEBASE as the knowledge base and conduct extensive experiments on the WEB QUESTIONS dataset. Experimental results show that our method achieves better or comparable performance compared with baseline systems. In addition, we develop a method to compute the salience scores of question words in different column networks. The results help us intuitively understand what MCCNNs learn.

7: A Statistical Model for Measuring Structural Similarity between Webpages

This paper presents a statistical model for measuring structural similarity between webpages from bilingual websites. Starting from basic assumptions we derive the model and propose an algorithm to estimate its parameters in unsupervised manner. Statistical approach appears to benefit the structural similarity measure: in the task of distinguishing parallel webpages from bilingual websites our language- independent model demonstrates an F- score of 0.94–0.99 which is comparable to the results of language-dependent methods involving content similarity measures.

8: CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text

We present the CLiPS Stylometry Investigation (CSI) corpus, a new Dutch corpus containing reviews and essays written by university students. It is designed to serve multiple purposes: detection of age, gender, authorship, personality, sentiment, deception, topic and genre. Another major advantage is its planned yearly expansion with each year's new students. The corpus currently contains about 305,000 tokens spread over 749 documents. The average review length is 128 tokens; the average essay length is 1126 tokens. The corpus will be made available on the CLiPS website (www.clips.uantwerpen.be/datasets) and can freely be used for academic research purposes. An initial deception detection experiment was performed on this data. Deception detection is the task of automatically classifying a text as being either truthful or deceptive, in our case by examining the writing style of the author. This task has never been investigated for Dutch before. We performed a supervised machine learning experiment using the SVM algorithm in a 10-fold cross-validation setup. The only features were the token unigrams present in the training data. Using this simple method, we reached a state-of-the-art F-score of 72.2

Schätze die Abstracts nach folgenden Gesichtspunkten ein:

1: Interessantheit

Verständlichkeit

Qualität

Originalität

2: Interessantheit

Verständlichkeit

Qualität

Originalität

3: Interessantheit

Verständlichkeit

Qualität

Originalität

4: Interessantheit

Verständlichkeit

Qualität

Originalität

5: Interessantheit

Verständlichkeit

Qualität

Originalität

6: Interessantheit

Verständlichkeit

Qualität

Originalität

7: Interessantheit

Verständlichkeit

Qualität

Originalität

8: Interessantheit

Verständlichkeit

Qualität

Originalität