LREC 2018 Workshop: MLP–MomenT
The Second Workshop on Multi-Language Processing in a Globalising World
The First Workshop on Multilingualism at the intersection of Knowledge Bases and Machine Translation

12 May 2018

ABSTRACTS

Editors:

Jinhua Du, Mihael Arcan, Qun Liu, Hitoshi Isahara

Workshop Programme

12 May 2018

09:00 – 09:10 Opening

09:10 – 09:50 Invited Talk 1: Prof. Toru Ishida, Kyoto University, Japan

09:50 – 10:30 **Session 1**: Multilingual Corpora and Lexical Similarity
09:50 – 10:10 Duygu Ataman, *Bianet: A Parallel News Corpus in Turkish, Kurdish and English*
10:10 – 10:30 Ohnmar Htuna, Koji Murakamib, Yu Hirate, *Phonetically Based Extraction of Japanese Synonyms from Rakuten Ichiba's Item Titles*

10:30 – 11:00 Coffee break

11:00 – 13:00 **Session 2**: Machine Translation and Multilingual NLP
11:00 – 11:20 Carlos Mullov, Jan Niehues, Alexander Waibel, *Inspection of Multilingual Neural Machine Translation*
11:20 – 11:40 Špela Vintar, *Terminology Translation Accuracy in Phrase-Based versus Neural MT: An Evaluation for the English-Slovene Language Pair*
11:40 – 12:00 Tao Feng, Miao Li, Lei Chen, *Low-Resource Neural Machine Translation with Transfer Learning*
12:00 – 12:20 Cristina Espana-Bonet, Josef van Genabith, *Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems*

12:20 – 13:00 Discussion

13:00 – 14:00 Lunch break

14:00 – 15:00 Invited Talk 2: Dr John P. McCrae, National University of Ireland Galway, Ireland

15:00 – 16:00 **Session 3**: Name Entity Recognition
15:00 – 15:20 Fei-Fei Liu, Zhi-Juan Wang, *Active Learning for Tibetan Named Entity Recognition based on CRF*
15:20 – 15:40 Zhijuan Wang, Fuxian Li, *A Semi-supervised Learning Approach for Person Name Recognition in Tibetan*

15:40 – 16:00 Discussion and End of the workshop

16:00 – 16:30 Coffee break

Workshop Organizers

Qun Liu	Dublin City University, Ireland
Hitoshi Isahara	Toyohashi University of Technology, Japan
Jinhua Du*	Dublin City University, Ireland
Mihael Arcan*	National University of Ireland Galway, Ireland
Tatjana Gornostaja	Tilde, Latvia, ELRA, BDVA
Darja Fišer	University of Ljubljana, Jožef Stefan Institute, CLARIN ERIC
Elena Montiel-Ponsoda	Universidad Politécnica de Madrid, Spain
Chao-Hong Liu	Dublin City University, Ireland
Joachim Wagner	Dublin City University, Ireland
Mikel L. Forcada	Universitat d'Alacant, Spain
Qi Zhang	Dublin City University, Ireland

*: Main editors and chairs of the Organising Committee

Workshop Programme Committee

Guadalupe Aguado-de-Cea	Universidad Politécnica de Madrid, Spain
Paul Buitelaar	National University of Ireland Galway, Ireland
Philipp Cimiano	Bielefeld University, Germany
Thatsanee Chaeronporn	Burapha University, Thailand
Christian Chiarcos	Goethe-Universität, Germany
Christopher Crowhurst	United Language Group, USA
Beatrice Daille	University of Nantes, France
Brian Davis	Maynooth University, Ireland
Thierry Declerck	German Research Center for Artificial Intelligence (DFKI), Germany
Mauro Dragoni	Fondazione Bruno Kessler (FBK), Italy
Tomaž Erjavec	Jožef Stefan Institute, Slovenia
Miquel Esplà-Gomis	Universitat d'Alacant, Spain
Natalia Grabar	Université de Lille, France
Jorge Gracia	University of Zaragoza, Spain
Miloš Jakubíček	University in Brno / Lexical Computing Limited, Czech
John Judge	Dublin City University, Ireland
Kyoko Kanzaki	Toyohashi University of Technology, Japan
Ilan Kernerman	K Dictionaries, Israel
Simon Krek	Jožef Stefan Institute, Slovenia
Els Lefever	Ghent University, Belgium
Qing Ma	Ryukoku University, Japan
Gudrun Magnusdottir	ESTeam AB, Sweden / Coreon GmbH, Germany
John Philip McCrae	National University of Ireland Galway, Ireland
Yohei Murakami	Ritsumeikan University, Japan
Roberto Navigli	Sapienza University of Rome, Italy
Mārcis Pinnis	Tilde, Latvia
Laurette Pretorius	University of South Africa (UNISA), South Africa
Gema Ramírez	Prompsit Language Engineering, Spain
Georg Rehm	German Research Center for Artificial Intelligence (DFKI), Germany
Virach Sornlertlamvanich	Sirindhorn International Institute of Technology (SIIT), Thailand
Antonio Toral	University of Groningen, Netherlands
Masatoshi Tsuchiya	Toyohashi University of Technology, Japan
Marco Turchi	Fondazione Bruno Kessler (FBK), Italy
Špela Vintar	University of Ljubljana, Slovenia
Eiko Yamamoto	Gifu Shotoku Gakuen University, Japan

Session 1: Multilingual Corpora and Lexical Similarity

Saturday 12 May, 09:50 – 10:30
Chairperson: TBD

Bianet: A Parallel News Corpus in Turkish, Kurdish and English

Author: Duygu Ataman

Abstract:

We present a new open-source parallel corpus consisting of news articles collected from the Bianet magazine, an online newspaper that publishes Turkish news, often along with their translations in English and Kurdish. In this paper, we describe the collection process of the corpus and its statistical properties. We validate the benefit of using the Bianet corpus by evaluating bilingual and multilingual neural machine translation models in English-Turkish and English-Kurdish directions.

Phonetically Based Extraction of Japanese Synonyms from Rakuten Ichiba's Item Titles

Authors: Ohnmar Htun, Koji Murakami, Yu Hirate

Abstract:

This paper presents a method for the phonetically based extraction of Japanese synonyms from item titles of Rakuten Ichiba. In general, synonyms are words with the same or similar meaning in a semantic sense; however, we focus here on those synonyms which appear as transliterations between English and Japanese, using Katakana, Hiragana, Kanji and a mixture of these scripts. The method consists of three parts: generation of the candidate word pairs using phrase detection (collocation) at the preprocessing stage; mapping similar sounds using Soundex and a cross-language sound group; measuring the similarity based on the Levenshtein and stochastic distances; and ranking the synonym pairs using fuzzy matching in the post-processing stage. We carry out two experiments based on two different sound mapping datasets, each of which measures the similarity scores from two different algorithms. The results from the baseline and cross-language models achieve precision values of 0.9208 and 0.9983, respectively. Our method is applicable to various fields of linguistic research, for example building a thesaurus/new name entity lookup for a search engine, machine translation and natural language generation, and improving output of voice recognition systems.

Session 2: Machine Translation and Multilingual NLP

Saturday 12 May, 11:00 – 12:20
Chairperson: TBD

Inspection of Multilingual Neural Machine Translation

Authors: Carlos Mullov, Jan Niehues, Alexander Waibel

Abstract:
In this paper we inspect the intermediate sentence representation in the multilingual attention-based NMT system proposed by Ha et al. (2016). We ask the question of how well the NMT system learns a shared representation across multiple languages, as such a shared representation is an important prerequisite for zero-shot translation. To this end we examine whether the sentence representation is inde- pendent of the individual languages involved in translation. Having found the sentence representation in our multilingual NMT system to be language dependent, we further inspect the sentence representation for the cause of this dependence. We isolated the language dependent features, and found present a linear correlation between the sentence representation and its source language. Using these isolated features, we describe a method to manipulate these features, and provide a way to eliminate the language specific differences between the sentence representations. This could potentially help to remove noise, which is particularly harmful for zero-shot translation.

Terminology Translation Accuracy in Phrase-Based versus Neural MT: An Evaluation for the English-Slovene Language Pair

Author: Špela Vintar

Abstract:
For specialised texts, the accuracy and consistency of terminology is of primary importance, yet most Machine Translation systems do not employ explicit strategies to ensure term consistency on the level beyond a single sentence. We present a multifaceted evaluation and comparison of a statistical phrase-based versus neural model of Google's translation system for the English-Slovene language pair, which consists of a document-based automatic evaluation with the BLEU and NIST metrics, an automatic evaluation of term translations using an existing termbase as reference, and a human evaluation of 300 sample sentences per MT model and translation direction. Results indicate that while neural MT regularly outperforms phrase-based MT in the overall scores, the accuracy of term translations is better only for the English-Slovene language pair and not in the Slovene-English translations. In the final part of the paper we discuss typical errors encountered in the different MT outputs.

Low-Resource Neural Machine Translation with Transfer Learning

Author: Tao Feng, Miao Li, Lei Chen

Abstract:
Neural machine translation has achieved great success under a great deal of bilingual corpora in the past few years. However, it does not work well for low-resource language pairs. In order to solve this problem, we present a transfer learning method which can improve the BLEU scores of the low-resource machine translation. First, we exploit encoder-decoder framework with attention mechanism to train one neural machine translation model with large language pairs, and then employ some parameters of the trained model to initialize another neural machine translation model with less bilingual parallel corpora. Our experiments demonstrate that the proposed method can achieve the excellent performance on low-resource machine translation by weight adjustment and retraining. On the IWSLT2015 Vietnamese-English translation task, our model can improve the translation quality by an average of 1.55 BLEU scores. Besides, we can also get the increase of 0.99 BLEU scores when translating from Mongolian to Chinese. Finally, we analyze the results of experiments and summarize our contribution.

Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems

Author: Cristina Espana-Bonet and Josef van Genabith

Abstract:
Neural machine translation systems are state-of-the-art for most language pairs despite the fact that they are relatively recent and that because of this there is likely room for even further improvements. Here, we explore whether, and if so, to what extent, semantic networks can help improve NMT. In particular, we (i) study the contribution of the nodes of the semantic network, synsets, as factors in multilingual neural translation engines. We show that they improve a state-of-the-art baseline and that they facilitate the translation from languages that have not been seen at all in training (beyond zero-shot translation). Taking this idea to an extreme, we (ii) use synsets as the basic unit to encode the input and turn the source language into a data-driven interlingual language. This transformation boosts the performance of the neural system for unseen languages achieving an improvement of 4.9/6.3 and 8.2/8.7 points of BLEU/METEOR for fr2en and es2en respectively when neither corpora in fr or es has been used. In (i), the enhancement comes about because cross-language synsets help to cluster words by semantics irrespective of their language and to map the unknown words of a new language into the multilingual clusters. In (ii), because with the data-driven interlingua there is no unknown language if it is covered by the semantic network. However, non-content words are not represented in the semantic network, and a higher level of abstraction is still needed in order to go a step further and train these systems with only monolingual corpora for example.

Session 3: Name Entity Recognition

Saturday 12 May, 15:00 – 15:40
Chairperson: TBD

Active Learning for Tibetan Named Entity Recognition based on CRF

Authors: Fei-Fei Liu, Zhi-Juan Wang

Abstract:
Named entity recognition (NER) is a major subtask of information extraction. Previous research tent to use huge amount of labeled data to train a classifier. But it is expensive for low resource languages One of the dominant problems facing Tibetan named entity recognition is the lack of training data. Active learning is a supervised machine learning algorithm which can achieve greater accuracy with fewer training labels. Active learning has been successfully applied to a number of natural language processing tasks, such as, information extraction, named entity recognition, text categorization, part-of-speech tagging, parsing, and word sense disambiguation. In this paper, we apply active learning based on Conditional Random Field (CRF) for Tibetan named entity recognition to minimize labeling effort by selecting the most informative instances to label. This paper proposes two kinds of query strategies, including Confidence, and Named Entity features. We compare the query strategies with the random method, and show that considerable performance improvements in reduce the human effort.

A Semi-supervised Learning Approach for Person Name Recognition in Tibetan

Authors: Zhijuan Wang, Fuxian Li

Abstract:
Massive labelled data is important for Named Entity Recognition(NER). For Low Resource Languages(LRL), massive labelled data means more labor, more time and more cost. A semi-supervised learning (SSL) that need fewer labelled data is proposed to recognize person name in Tibetan texts. Based on Conditional Random Fields (CRFs) and Radial Basis Function (RBF), this method use 5-element feature matrix to propagate information from few labeled data to massive unlabelled data. Experiments demonstrate that its F-measure can achieve 84% using only 100 documents as seeds, whereas about 800 labeled documents are required for a supervised learning based on pure CRFs.