BSNLP2019 Shared Task Submission:
Multisource Neural NER Transfer

Tatiana Tsygankova, Stephen Mayhew, Dan Roth
University of Pennsylvania
Philadelphia, PA, 19104
{ttasya, mayhew, danroth}@seas.upenn.edu

Abstract

This paper describes the Cognitive Computation (CogComp) Group's submissions to the multilingual named entity recognition shared task at the Balto-Slavic Natural Language Processing (BSNLP) Workshop (Piskorski et al., 2019). The final model submitted is a multi-source neural NER system with multilingual BERT embeddings, trained on the concatenation of training data in various Slavic languages (as well as English). The performance of our system on the official testing data suggests that multi-source approaches consistently outperform single-source approaches for this task, even with the noise of mismatching tagsets.

1 Introduction

This paper describes the Cognitive Computation (CogComp) Group's submission to the shared task of the Balto-Slavic Natural Language Processing (BSNLP) Workshop at ACL 2019 (Piskorski et al., 2019). This shared task centers around multilingual named entity recognition (NER) in Slavic languages, and is composed of recognition, lemmatization, and entity linking subtasks. The niche focus of this task on Slavic languages makes it both interesting and challenging. The languages used in the shared task (Bulgarian, Czech, Polish, and Russian) belong to the same language family and share complex grammatical and morphological features which may be understudied in an English-focused research community. Further, they encompass both Latin and Cyrillic scripts, complicating the multilingual nature of the problem. In addition to the language specific challenges, there are varying sizes of training data, somewhat non-standard named entity types (making finding additional data challenging), and differing domains – the training and test sets are composed of newswire documents collected around domain-specific topics, with different topics in train and test.

This year's shared task is the second edition of the multilingual named entity recognition task on Slavic languages organized for the BSNLP workshop. A similar shared task was previously held in 2017 (BSNLP2017), and was composed of the same subtasks, but was evaluated on seven Slavic languages. It had a slightly different format, in that training data was not provided to the participants, so the majority of the submissions relied on cross-lingual or rule-based approaches.

Our overarching research goal for this project was to experiment with multisource neural NER transfer, leveraging recent advances in multilingual contextual embeddings (Devlin et al., 2019). Ultimately, we aimed to maximize parameter-sharing by training a single model on the concatenation of training data from sources (languages). Such *multi-source* systems have seen success in machine translation (Zoph and Knight, 2016), and to some extent in non-neural NER systems (Mayhew et al., 2017), and neural systems (Rahimi et al., 2019). Given that training data is available in this iteration of the shared task, we purposefully chose to not include rule-based components into our model in order to focus on getting the most out of the given training data.

Our results on the official test data show that multi-source models using multilingual contextual embeddings produce strong performance, and incorporating a greater variety of languages within the same language family further boosts the results. We also observe that combining training data from distinct tagsets often improves performance, and generalizes to the intended tagset better than expected. Finally, our experiments using cross-lingual NER trained on English showed results inferior to monolingual experiments, but surprisingly high nonetheless.

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 75–82,
Florence, Italy, 2 August 2019. ©2019 Association for Computational Linguistics

2 Related Work

The first shared task in Balto-slavic NLP was held in 2017, and reported in Piskorski et al. (2017). The task was somewhat different from the 2019 task in that training data was not provided to participants. Approaches submitted to this task included a model based on parallel projection (Mayfield et al., 2017) and a model with language-specific features trained on found data (Marcińczuk et al., 2017). There has also been follow-up work on this dataset using cross-lingual embeddings (Sharoff, 2018).

Named Entity Recognition (NER), the task of detecting and classifying named entities in text, has been studied for many years. Early models proposed were averaged perceptron (Ratinov and Roth, 2009), and conditional random field (Manning et al., 2014). In recent years, neural models have proved successful, with the BiLSTM-CRF model dominant (Chiu and Nichols, 2016; Lample et al., 2016). A further increase in performance has come with contextual embeddings (Devlin et al., 2019; Peters et al., 2018; Akbik et al., 2018), which are based on large language models trained over massive corpora.

Of particular interest is the multilingual BERT model (Devlin et al., 2019), which is trained over the concatenation of the Wikipedias in over 100 languages.[1] Although BERT is not trained with explicit cross-lingual objectives, it has been shown to have emergent cross-lingual properties, as well as language identification capabilities (Wu and Dredze, 2019).

Several models have been proposed for multi-source learning, in which multiple languages are used to train a model, including for machine translation (Zoph and Knight, 2016; Johnson et al., 2017; Currey and Heafield, 2018), and NER (Täckström, 2012; Tsai et al., 2016; Mayhew et al., 2017; Rahimi et al., 2019).

3 Task

We first describe the details of the shared task, including the data, the evaluation metrics, and the subtasks.

3.1 Data

The BSNLP 2019 training set contained four Slavic languages: Bulgarian, Czech, Polish and Russian. Of these, Czech and Polish are written in Latin script, and Russian and Bulgarian are written in Cyrillic script, a property that we will later explore in our experiments. Table 1 summarizes the size of the datasets. There is a large disparity in the amount of training data by language, with the largest (Polish), containing almost 3 times as many tokens as the smallest (Russian). The training data is in the form of newswire articles and contains document-level annotations of five different entity types: persons (PER), locations (LOC), organizations (ORG), events (EVT) and products (PRO). In document-level supervision, the entity annotations are given for each document as a list of unique surface forms of entities and their corresponding tags, but with no span information. Although this is quite different from the token-level annotations used more commonly for NER data, we argue later that it's possible to convert between the two formats in a (mostly) lossless fashion.

Lang.	Docs	Tokens
English (CoNLL)	964	203,621
Bulgarian (BG)	699	226,728
Czech (CS)	373	84,636
Polish (PL)	586	237,333
Russian (RU)	271	67,495

Table 1: Training data sizes in CoNLL and BSNLP19 datasets. Of the BSNLP19 sets, the largest (Polish) is nearly 3 times the size of the smallest (Russian).

Tag	Total	Unique	Ratio
PER	9986	2851	3.5
LOC	9563	1540	6.2
ORG	8520	1923	4.4
EVT	2601	235	11.0
PRO	1699	739	2.3

Table 2: Entity distribution statistics across all languages in the BSNLP19 training set, where the "Ratio" column refers to the proportion of the "Total" number of entity type annotations to the "Unique" annotations.

The training documents are divided into two topics: one set containing news articles relating to Brexit, and the other with news articles about a Pakistani woman named Asia Bibi. These focused domains suggest that the set of unique entities will be relatively small within each topic. Table 2 supports this hypothesis and shows the distribution of total and unique entity tags for the entire training set. The high ratio of total to unique mentions for certain tags such as event (EVT) means that

[1] github.com/google-research/bert/blob/master/multilingual.md

the training data contains a small variety of distinct surface forms labeled as "EVT", which could lead to potential overfitting to these entities. Given that the test set used for evaluation of our models contains news articles surrounding two distinct topics (containing documents about Nord Stream, an offshore gas pipeline in Russia, and Ryanair, an Irish low-cost airline), it's also likely that the small number of unique entities could lead to poor domain generalization results for those tags.

3.2 Evaluation Metrics

Since the shared task annotations are created on the document level, the evaluation metrics are somewhat different from standard NER. They are similarly based on precision, recall, and F1 measure of retrieved entities, but are based on matching surface forms between sets of entities instead of matching spans. When matching surface forms, two types of evaluation are used. These are described in the official documentation[2] as:

- **Relaxed evaluation**: an entity mentioned in a given document is considered to be extracted correctly if the system response includes at least one annotation of a named mention of this entity (regardless whether the extracted mention is base form); This is evaluated in two ways:
  - **Partial match**: partial matches count.
  - **Exact match**: full string must match.
- **Strict evaluation**: the system response should include exactly one annotation for each unique form of a named mention of an entity that is referred to in a given document, i.e., capturing and listing all variants of an entity is required. There is no partial score given for this metric.

In our analysis below, we chose to report Strict evaluation as being the most similar to the span-based F1 commonly used in NER.

3.3 Subtasks

Within this shared task there are three distinct subtasks: Recognition, Lemmatization, and Linking. We focus only on the Recognition task.

4 Experimental Setup

4.1 Annotation Conversion

Given that the annotations for the training data were provided at the document-level, we decided to simply convert these to token-level annotations in order to use standard token-level NER tools. We performed the conversion by traversing each of the annotated entities in the list of document-level annotations and extrapolating the named entity tags to matching surface level forms in the original document. For example, if the list of document-level annotations for some document X contained an annotation for "Brexit" as event (EVT), we would tag all instances of "Brexit" in document X as event at the token-level, and assign "O" to everything that does not have an annotation.

This conversion is susceptible to two types of annotation errors: tagging a token as a named entity when it should be tagged as "O", and tagging a token as an incorrect named entity type.

Although we have no sure way of estimating the error from the first type aside from inspection, experience suggests that such situations are relatively rare in Slavic languages.[3] For example, an entity like *Nunzio Galantino* (a person) is virtually always a person.

As for the second type of error, we found that only 15 documents contained a surface form with multiple entity tags. We decided that this small number of errors is insignificant, and would add very little noise.

For the official evaluation, we made token-level predictions on the test data and converted them to the document level submission format.

4.2 Additional Data

In our experiments, we included two additional datasets – the testing data from the previous iteration of the BSNLP Multilingual NER Shared Task composed of document-level annotations for 7 Balto-Slavic languages, and the English CoNLL 2003 data. What made the use of these datasets challenging was that both were labeled with the CoNLL 2003 entity types – PER, LOC, ORG and MISC – a set not identical to that of the BSNLP19 data. In theory, such a mismatch would be prohibitive, since it would result in unwanted MISC tags, and missed EVT and PRO tags in our output. However, in our preliminary experiments, we were surprised to learn that tagset mismatch across languages seemed to not be a problem (see more discussion of this phenomenon in Section 6). Models trained on data with MISC tags occasionally pro-

[2] bsnlp.cs.helsinki.fi/shared_task.html

[3] This may be more of a problem in Semitic languages, for example.

duced MISC tags in the output (less than 10 times in the test data), but we simply removed these predictions at post-processing time.

We hypothesized that the model is able to associate language with tagset, and accordingly only used BSNLP 2017 languages that were not present in our training set, that is: Croatian, Slovak, Slovene, and Ukrainian.

4.3 Preliminary Experiments

In our preliminary experiments, we created a development set to measure the relative improvement of each idea. Given that our training set was composed of documents surrounding two distinct topics, our initial approach was to create a multitopic validation split, where the development set contained documents from both topics. However, our models reached nearly perfect scores on this split due to the small variation of entities within a given topic. This split was not representative of the official test set evaluation, since the testing data contains entirely new topics, and a lot more generalization would be needed. To better imitate testing conditions, we split the training data by topic, using one topic for training, and the other for development. Our preliminary experiments (not reported here) showed that using off-the-shelf multilingual FastText embeddings[4] (Joulin et al., 2018) resulted in significantly worse performance than BERT, and so omitted them from our submissions.

4.4 Model

For our model, we use a standard BiLSTM-CRF (Lample et al., 2016) implemented in AllenNLP (Gardner et al., 2018). The model used character embeddings with a single layer Convolutional Neural Network (CNN) with 128 filters, and word embeddings from multilingual BERT (Devlin et al., 2019). We used the bert-base-multilingual-cased model from huggingface[5] which uses a shared wordpiece vocabulary among all languages, meaning that we can share models even across Cyrillic and Latin scripts. We did not fine-tune BERT during training, but learned a scalar mix of the 12 layers. For each word, we use the first wordpiece to be representative of the entire word, as done in Devlin et al. (2019).

5 Results

Our main results are shown in Table 3, as F1 scores from the Strict evaluation (results from all metrics can be seen in the Appendix). We made a total of 8 submissions to the shared task, with each row in the table denoting a separate submission, with the exception of the first 4 rows. Those together composed one submission, since we tested each single-source model only on the same target language. Each submission in the table is also given a name (e.g. *LatinScript*) that is descriptive of the training data used. The columns are divided into two sections: training data on the left, and testing data on the right, both separated into various languages. The checkmarks denote which datasets were included in training. The rows of the table are divided into two sections, with the upper section representing single-source systems (using only one language in training), and the lower section representing multi-source models.

BSNLP17 training corpus refers to the testing data from the BSNLP shared task in 2017, as described in Section 4.2. **EN** refers to the CoNLL 2003 English training set.

6 Analysis

There are several interesting lessons in our results. First, multi-source training with BERT is a success, as evidenced by the 2.7 F1 improvement between the single-source experiments and the best experiment (*AllLangs*).

Surprisingly, these results hold even in the face of tagset mismatches. Recall from Section 4.2 that English CoNLL (EN) and the BSNLP17 datasets use a tagset somewhat different from the BSNLP19 test data. Despite this, we see an overall improvement from *AllTrain* (which does not use additional data from the BSNLP17 languages) to *AllLangs* (which does), and similarly from *AllTrainEng* to *AllLangsEng*. We believe that two factors contributed to this success:

Factor 1. The large overlap in the tagset distributions. PER, LOC, and ORG tags made up the majority of annotations in all datasets. Thus, most information required to learn a model is present in the training data regardless of tagset. Furthermore, PRO and EVT entities are rare enough in the test data that even small scores shouldn't hurt the micro-average. In fact, Table 4 shows that when going from *AllTrain*, which uses only the BSNLP19 tagset, to *AllLangsEng*, which includes

[4] fasttext.cc/docs/en/aligned-vectors.html

[5] github.com/huggingface/pytorch-pretrained-BERT

		Training Data						Testing Data				
	Submission	**BG**	**CS**	**PL**	**RU**	**EN**	**BSNLP17**	**BG**	**CS**	**PL**	**RU**	**ALL**
Single-source	Bulgarian	✓						80.9	–	–	–	82.0
	Czech		✓					–	84.0	–	–	82.0
	Polish			✓				–	–	85.2	–	82.0
	Russian				✓			–	–	–	76.8	82.0
	English					✓		74.3	76.0	72.2	73.3	73.9
Multi-source	LatinScript		✓	✓				78.1	87.8	85.2	77.0	82.6
	LatinScriptEng		✓	✓		✓		77.9	87.8	85.6	77.2	82.8
	AllTrain	✓	✓	✓	✓			82.7	88.0	85.9	**79.4**	84.3
	AllTrainEng	✓	✓	✓	✓	✓		82.8	87.8	85.6	78.5	84.0
	AllLangs	✓	✓	✓	✓		✓	**84.1**	88.3	86.1	79.3	**84.7**
	AllLangsEng	✓	✓	✓	✓	✓	✓	83.0	**88.5**	**86.3**	78.3	84.4

Table 3: Official results on the Recognition task of BSNLP19, measured as F1 with Strict evaluation. The training languages used are: Bulgarian (BG), Czech (CS), Polish (PL), Russian (RU), English (EN, CoNLL2003) and the BSNLP17 languages (Croatian, Slovak, Slovene and Ukrainian). The top section of the table shows single-source experiments, in which each model is trained on a single language. The bottom section shows multi-source experiments. The rightmost column, **ALL**, is a micro-average of the test results over the 4 test languages.

Method	PER	LOC	ORG	PRO	EVT
P. AllTrain	90.9	94.1	90.6	77.6	48.1
P. AllLangsEng	92.3	95.1	90.9	75.2	32.7
R. AllTrain	94.2	97.8	89.2	54.2	27.8
R. AllLangsEng	95.8	97.7	86.8	60.6	31.5
F1. AllTrain	92.5	95.9	89.9	63.9	35.3
F1. AllLangsEng	94.0	96.3	88.9	67.1	32.1

Table 4: Precision (P), Recall (R), and F1 scores by tag across all languages. *AllTrain* is the largest set of training data that uses solely the target tagset, and *AllLangsEng* includes training data with the tagset with MISC and without PRO or EVT.

data with the divergent tagset, the recall on EVT and PRO actually improves.

Factor 2. The power of multilingual BERT. We know that multilingual BERT can detect language (Wu and Dredze, 2019), and we hypothesize that multilingual BERT is able to associate language with tagset.

While we show that multi-source training data helps, our results also show that choosing the right languages for inclusion is important. Naturally, scores are better if the target language is present in the training data, with the exception of Single-source Russian compared with *LatinScript* Russian. This could be attributed to the fact that there is relatively little Russian training data, and the model is powerful enough that a large amount of Polish and Czech data is better than a small amount Russian data. Even so, scores further improve when Russian is added again (*AllTrain*).

Finally, there are some interesting observations on the model trained only on English data. It performs well both across tagsets, and across scripts (on Bulgarian and Russian). Although one might expect that this approach would perform best on Latin script languages, such a correlation is not present. Further, scores across languages are within 4 points of each other, compared to individual monolingual systems that range over 10 points.

7 Conclusion

This paper has described our submission the BSNLP19 shared task on named entity recognition. Our approach is based on multi-source neural NER transfer, with experiments contrasting single-source and cross-lingual approaches. We found that using more data almost always helps, at least when in the same family.

Acknowledgement

This work was supported by Contracts HR0011-15-2-0025 and HR0011-18-2-0052 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Anna Currey and Kenneth Heafield. 2018. Multi-source syntactic neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2961–2966, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Michał Marcińczuk, Jan Kocoń, and Marcin Oleksy. 2017. Liner2 — a generic framework for named entity recognition. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 86–91, Valencia, Spain. Association for Computational Linguistics.

James Mayfield, Paul McNamee, and Cash Costello. 2017. Language-independent named entity analysis using parallel projection and rule-based disambiguation. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 92–96, Valencia, Spain. Association for Computational Linguistics.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The second cross-lingual challenge on recognition, classification, lemmatization, and linking of named entities across Slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The first cross-lingual challenge on recognition, normalization, and matching of named entities in Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 76–85, Valencia, Spain. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Serge Sharoff. 2018. Language adaptation experiments via cross-lingual embeddings for related languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Oscar Täckström. 2012. Nudging the envelope of direct transfer methods for multilingual named entity recognition. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 55–63, Montréal, Canada. Association for Computational Linguistics.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228, Berlin, Germany. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. *arXiv preprint arXiv:1904.09077*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

A Detailed Results

The full summary of our results with all submissions and all evaluation metrics is shown in Table 6. Table 1 has the key that maps between submission ID and name used in the main paper.

Model	**Submission Key**
Bulgarian	ccg-2
Czech	ccg-2
Polish	ccg-2
Russian	ccg-2
English	ccg-8
LatinScript	ccg-3
LatinScriptEng	ccg-4
AllTrain	ccg-1
AllTrainEng	ccg-5
AllLangs	ccg-6
AllLangsEng	ccg-7

Table 5: Key matching the descriptive submission names used throughout the paper with the submission numbers referenced in our results section.

All Corpora	*Language*							
Metric	bg		cs		pl		ru	
Relaxed Partial	ccg-1	86.9	ccg-1	93.5	ccg-1	92.1	ccg-1	**88.6**
	ccg-2	85.1	ccg-2	92.0	ccg-2	92.0	ccg-2	86.0
	ccg-3	84.3	ccg-3	93.2	ccg-3	91.9	ccg-3	88.0
	ccg-4	84.3	ccg-4	93.6	ccg-4	**92.4**	ccg-4	87.7
	ccg-5	88.1	ccg-5	93.5	ccg-5	91.9	ccg-5	88.0
	ccg-6	**88.9**	ccg-6	93.5	ccg-6	92.0	ccg-6	88.5
	ccg-7	87.6	ccg-7	**94.0**	ccg-7	92.3	ccg-7	88.3
	ccg-8	81.0	ccg-8	83.4	ccg-8	78.6	ccg-8	83.5
Relaxed Exact	ccg-1	83.8	ccg-1	87.3	ccg-1	85.0	ccg-1	**81.4**
	ccg-2	82.0	ccg-2	83.2	ccg-2	84.3	ccg-2	78.3
	ccg-3	79.1	ccg-3	87.0	ccg-3	84.1	ccg-3	78.2
	ccg-4	78.7	ccg-4	87.0	ccg-4	84.7	ccg-4	78.3
	ccg-5	84.0	ccg-5	87.0	ccg-5	84.7	ccg-5	80.1
	ccg-6	**85.3**	ccg-6	87.9	ccg-6	**85.4**	ccg-6	81.0
	ccg-7	84.0	ccg-7	**88.0**	ccg-7	**85.4**	ccg-7	80.4
	ccg-8	75.5	ccg-8	74.5	ccg-8	70.1	ccg-8	74.1
Strict	ccg-1	82.7	ccg-1	88.0	ccg-1	85.9	ccg-1	**79.4**
	ccg-2	80.9	ccg-2	84.0	ccg-2	85.2	ccg-2	76.8
	ccg-3	78.1	ccg-3	87.8	ccg-3	85.2	ccg-3	77.0
	ccg-4	77.9	ccg-4	87.8	ccg-4	85.6	ccg-4	77.2
	ccg-5	82.8	ccg-5	87.8	ccg-5	85.6	ccg-5	78.5
	ccg-6	**84.1**	ccg-6	88.3	ccg-6	86.1	ccg-6	79.3
	ccg-7	83.0	ccg-7	**88.5**	ccg-7	**86.3**	ccg-7	78.3
	ccg-8	74.3	ccg-8	76.0	ccg-8	72.2	ccg-8	73.3

Table 6: Evaluation results (topics combined)