Proceedings of the 13th Workshop on Building and Using Comparable Corpora, pages 6–13
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Overview of the Fourth BUCC Shared Task: Bilingual Dictionary Induction from Comparable Corpora

Reinhard Rapp
Athena R.C. and Magdeburg-Stendal
University of Applied Sciences
and University of Mainz
reinhardrapp@gmx.de

Pierre Zweigenbaum
Université Paris-Saclay
LIMSI, CNRS
91400 Orsay, France
pz@limsi.fr

Serge Sharoff
University of Leeds
Leeds, LS2 9JT
United Kingdom
s.sharoff@leeds.ac.uk

Abstract

The shared task of the 13th Workshop on Building and Using Comparable Corpora was devoted to the induction of bilingual dictionaries from comparable rather than parallel corpora. In this task, for a number of language pairs involving Chinese, English, French, German, Russian and Spanish, the participants were asked to determine automatically the target language translations of several thousand source language test words in three frequency ranges. We describe here some background, the task definition, the training and test data sets and the evaluation used for ranking the participating systems. We also summarize the approaches used and present the results of the evaluation. In conclusion, the outcome of the competition is the results of a number of systems which provide surprisingly good solutions to an ambitious problem.

Keywords: bilingual dictionary, lexicon induction, comparable corpora

1. Introduction

In the framework of machine translation, the extraction of bilingual dictionaries from parallel corpora has been conducted very successfully (see e.g. Mihalcea & Pedersen, 2003). But on the other hand, human second language acquisition appears not to be based on parallel data. This means that there must be a way of acquiring and relating lexical knowledge across two or more languages without the use of parallel data.

It has been suggested that it may be possible to extract multilingual lexical knowledge from comparable rather than from parallel corpora (see e.g. Sharoff et al., 2013). From a theoretical perspective, this suggestion may lead to advances in understanding human second language acquisition. From a practical perspective, as comparable corpora are available in much larger quantities than parallel corpora, this approach might help in relieving the data acquisition bottleneck which tends to be especially severe when dealing with language pairs involving low resource languages (see e.g. Martin et al., 2005).

A well-established practical task to approach this topic is bilingual lexicon extraction from comparable corpora, which is in the focus of this shared task. Typically, its aim is to extract word translations such as exemplified in Table 1 from comparable corpora, where a given source word may receive multiple translations. Note that, to reflect the tabular format used in the shared task, multiple translations of the same source word are listed in separate rows.

Quite a few research groups have been working on this problem using a wide variety of approaches. There are comprehensive studies such as Irvine & Callison-Burch (2017) and also overview papers at least in part discussing the topic like Jakubina & Langlais (2016), Rapp et al. (2016), Sharoff et al. (2013).

Source (English)	Target (French)
baby	bébé
baby	poupon
bath	bain
bed	lit
bed	plumard
convenience	commodité
doctor	médecin
doctor	docteur
eagle	aigle
mountain	montagne
nervous	nerveux
work	travail

Table 1: Sample word translations from English to French. In the shared task a similar tab-separated format was used.

However, as up to now there was no standard way to measure the performance of the systems, the published results are not comparable and the pros and cons of the various approaches are not clear.

2. Shared Task Description

The present shared task[1] aimed at solving these problems by organizing a fair competition between systems. This was accomplished by providing corpora and bilingual datasets for a number of language pairs involving Chinese, English, French, German, Russian and Spanish, and by comparing the results using a common evaluation framework. For the shared task we provided corpora as well as training and test data. However, as we anticipated that these corpora and datasets may not suit all needs, we divided the shared task into two tracks:

- In the *closed track*, participants were required to only use the data provided by the organizers. In this way equal conditions were ensured and, as the outcome of

[1] https://comparable.limsi.fr/bucc2020/bucc2020-task.html

this track, the systems could be compared and ranked according to the quality of their results.

- In the *open track*, participants were free to use their own corpora and training data. If possible, they were supposed to still use the evaluation data provided in the closed track, but this was also not mandatory. The participants could even work on languages for which the shared task provided no data. If relevant, the participants were supposed to describe why their systems were not suitable for the closed track, and discuss the pros and cons of their choices. They were also encouraged to provide access to their data for the purpose of facilitating replication by others.

To give an overview on the steps to be conducted by the participating teams, Table 2 provides a checklist for the participants in an abbreviated form. The time schedule is shown in Table 3. With about three weeks, the time span between the release of the test sets and the submission of the final results was (in comparison to most other shared tasks) foreseen to be relatively long for the reason that some teams worked on more language pairs than others and would have been at a disadvantage if this time span had been a limiting factor (but it probably still was to some extent).

- Decide on the track and the language pairs.
- Express your interest to the shared task organizers. You may also suggest new language pairs, and we might be able to help you with data.
- Download the corpora from the shared task webpage (WaCky or Wikipedia)
- Download the training data (bilingual word pairs) for your language pairs from the shared task webpage.
- Run your system on the words on the source side of the training data and compute the translations. Compare your results with target side of the training data and improve your system if necessary.
- Download the test data on the date specified in the time schedule.
- Run your system on the test data. Format your output in the same way as you see in the training data.
- Before the deadline specified in the schedule, submit your results.
- Write and submit a system description paper.
- Present your paper at the workshop.

Table 2: Checklist for participants (abbreviated).

Any time	Expressions of interest to participate in the shared task
January 12, 2020	Release of shared task training sets
February 16, 2020	Release of shared task test sets
March 5, 2020	Submission of shared task results

Table 3: Time schedule.

3. Closed Track

3.1 Corpora

Table 4 lists the corpora to be used for the language pairs supported in the closed track. Due to their free availability for several languages and their size, for the shared task we used the WaCky-corpora kindly provided by the Web-as-a-corpus initiative[2] (Baroni et al., 2009) and cleaned-up versions of Wikipedia dumps.

The cells in Table 4 show which of the two types of corpora were supposed to be used for the two languages of a language pair when conducting the dictionary induction task. The rationale behind these choices is that the WaCky corpora, with a greater variety of topics and genres, seem somewhat better suited for the dictionary induction task than Wikipedia, but they are not available for Chinese and Spanish. Language pairs involving Chinese and Spanish therefore use Wikipedia, whereas other language pairs use WaCky.

	de	**en**	**es**	**fr**	**ru**	**zh**
de		deWaC ukWaC	deWiki esWiki	deWaC frWaC		
en	ukWaC deWaC		enWiki esWiki	ukWaC frWaC	ukWaC ruWaC	enWiki zhWiki
es	esWiki deWiki	esWiki enWiki				
fr	frWaC deWaC	frWaC ukWaC				
ru		ruWaC ukWaC				
zh		zhWiki enWiki				

Table 4: Language pairs supported and corpora (WaCky or Wikipedia) to be used in the closed track.

The WaCky corpora are cleaned-up web crawls. Their compressed sizes are: English: 3.2 GB, French: 3.0 GB, German; 3.0 GB, Russian: 4.1 GB. English, French, and German are supposed to comprise in the order of 2 billion, Russian about 3 billion running words (Sharoff et al., 2017).

The compressed sizes of the Wikipedia corpora are: English: 3.6 GB, Spanish: 0.9 GB, Chinese: 0.4 GB. They are in a one-line per document format. The first tab-separated field in each line contains metadata, the second field contains the text. Paragraph boundaries are marked with HTML tags. As cleaning up the original Wikipedia dump files is not trivial, occasionally there can be some noise in the form of not fully cleaned HTML and Javascript fragments. Details of the cleanup and preparation procedure can be found in Sharoff et al. (2015).

3.2 Embeddings

For the convenience of the shared task participants, we provided pre-trained fastText embeddings for all WaCky and Wikipedia corpora listed in Table 4. They were trained

[2] http://wacky.sslmit.unibo.it/doku.php?id=corpora

on the Wikipedia or WaCky corpora and were allowed to be readily used in both tracks.

The fastText embeddings for the Wikipedia corpora were taken from Facebook AI Research (Bojanowsky et al., 2017).[3] For the WaCky-corpora, pre-trained fastText embeddings were computed and made available by Serge Sharoff as follows:

- The .vec.xz files are text representations, widely used in various tools.
- The .bin files are binary versions for use in fastText.
- The following parameters were used: method: skipgram; minCount: 30; dim: 300; ws (context window): 7; epochs: 10; neg (number of negatives sampled): 10. The other parameters are the defaults for fastText.

3.3 Training and test datasets

For training and testing the systems, reasonable numbers of bilingual word pairs as exemplified in Table 1 had to be provided for the language pairs listed in Table 4. Alexis Conneau from Facebook AI Research kindly gave us permission to use for the shared task extracts from the MUSE "Ground-truth bilingual dictionaries"[4] as described in Conneau et al. (2017). In this paper, the authors describe their data as follows:

> "**Word translation** The task considers the problem of retrieving the translation of given source words. The problem with most available bilingual dictionaries is that they are generated using online tools like Google Translate, and do not take into account the polysemy of words. Failing to capture word polysemy in the vocabulary leads to a wrong evaluation of the quality of the word embedding space. Other dictionaries are generated using phrase tables of machine translation systems, but they are very noisy or trained on relatively small parallel corpora. For this task, we create high-quality dictionaries of up to 100k pairs of words using an internal translation tool to alleviate this issue. We make these dictionaries publicly available as part of the MUSE library"

To us, the MUSE data on word translations looks like being derived from word-aligned parallel corpora by filtering out infrequent and therefore less reliable translations of a source language word. In particular, as it seems that for each source language word at most five possible translations are provided, it appears that only those target language translations which are aligned to at least 20% of the occurrences of a given source language word are listed.[5]

For more than 100 language pairs, the MUSE data lists such word translations. The lists use UTF-8 encoding and lower case characters only. Apparently, they are sorted by descending corpus frequencies of the source language words. As an example, Table 5 shows the top 40 lines of the list for English–German. For some language pairs, blanks are used as separators between source word and translation, but tabs for others. Although this is not applicable to the current shared task, to provide for future extensions to multiword units, we unified this to tabs.

English	German	English	German
the	die	utc	utc
the	der	his	seinem
the	dem	his	seinen
the	den	his	seine
the	das	his	sein
and	sowie	his	seiner
and	und	not	not
was	war	not	nicht
was	wurde	not	kein
for	für	are	sind
that	dass	talk	vortrag
that	das	talk	gespräch
with	mit	talk	reden
from	vom	talk	talk
from	von	which	welches
from	ab	which	welcher
from	aus	which	welche
this	dieser	which	welchen
this	diese	also	ausserdem
this	das	also	ebenso

Table 5: Top 40 translations from the English to German MUSE word translation data.

Table 6 gives, in alphabetical order according to ISO-language codes,[6] an overview of the number of bilingual word pairs (lines in the files) provided for each of the language pairs in the MUSE word translation data.[7] As can be seen in column *Lines*, this number varies between 20549 (ko-en) and 113324 (fr-en). However, as many source language words have several translations, the number of unique source language words (word types) is smaller. Column *Types* shows that this number varies between 13727 (ko-en) and 106473 (es-pt). Comparing the two columns gives an idea of the average number of translations for each source language word of a language pair.

Rather than providing one large set of training data for each language pair, by splitting into three frequency ranges we provide three equally-sized smaller sets per language pairs. Looking at different frequency ranges is of scientific interest as algorithms typically work best for high or medium frequency words, whereas the performance at low frequencies is often of higher practical relevance.

[3] https://fasttext.cc/docs/en/pretrained-vectors.html

[4] https://github.com/facebookresearch/MUSE

[5] We are extrapolating from what we did ourselves in the previous COMTRANS project, which, however, covered only a few language pairs (https://cordis.europa.eu/project/id/23845)

[6] https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

[7] As of May 2020, the MUSE website lists dictionaries for 110 language pairs (see https://github.com/facebookresearch/MUSE). However, there is a double occurrence of the en-en file (identical files with the same English words on the source and the target side). We list this file only once in our table which is why we have only 109 items in Table 6.

Lang.	Lines	Types	Lang.	Lines	Types
af-en	37421	36054	en-tr	67799	58901
ar-en	31355	24547	en-uk	47912	39365
bg-en	55170	45769	en-vi	77020	74447
bn-en	23829	19165	en-zh	39334	32495
bs-en	43318	40997	es-de	68869	59839
ca-en	78081	76720	es-en	112583	96579
cs-en	64211	55867	es-fr	87297	86765
da-en	81959	70776	es-it	96290	95406
de-en	101997	78200	es-pt	107363	106473
de-es	68905	64574	et-en	32776	28527
de-fr	61527	60802	fa-en	41321	33914
de-it	59811	59373	fi-en	43102	35770
de-pt	54765	54554	fr-de	61517	61119
el-en	45515	37186	fr-en	113324	97021
en-af	37446	35000	fr-es	87310	87010
en-ar	33663	22305	fr-it	97719	97121
en-bg	61240	49447	fr-pt	94552	92193
en-bn	30737	25564	he-en	45679	36735
en-bs	43333	38784	hi-en	31046	25732
en-ca	78097	74867	hr-en	56424	51305
en-cs	64216	52554	hu-en	42823	34974
en-da	82018	67177	id-en	96518	83355
en-de	101931	74655	it-de	59798	59686
en-el	56070	45152	it-en	103613	93214
en-en	92844	92844	it-es	96284	91929
en-es	112580	93084	it-fr	97711	92706
en-et	32748	27514	it-pt	91869	91503
en-fa	48164	41327	ja-en	25969	21003
en-fi	43055	32061	ko-en	20549	13727
en-fr	113286	94681	lt-en	33435	31807
en-he	47333	38070	lv-en	46385	40419
en-hi	38221	31719	mk-en	43935	36620
en-hr	56413	47834	ms-en	73092	66469
en-hu	42868	34944	nl-en	93853	84583
en-id	96500	86656	no-en	75171	70035
en-it	103612	90589	pl-en	73901	64397
en-ja	35353	31580	pt-de	54737	54534
en-ko	22357	17517	pt-en	108686	97261
en-lt	33447	30595	pt-es	107351	102289
en-lv	46407	37250	pt-fr	94517	88109
en-mk	50749	40580	pt-it	91849	91370
en-ms	73087	68548	ro-en	80821	75407
en-nl	93835	82181	ru-en	48714	40486
en-no	75204	66098	sk-en	65878	56408
en-pl	73883	59952	sl-en	62890	54894
en-pt	108696	92504	sq-en	52090	45534
en-ro	80815	68749	sv-en	82348	71678
en-ru	53186	42615	ta-en	21230	18247
en-sk	65887	50917	th-en	25332	21567
en-sl	62907	51473	tl-en	34984	32284
en-sq	52111	40853	tr-en	68611	58040
en-sv	82372	68608	uk-en	40723	34888
en-ta	26656	22610	vi-en	76364	73445
en-th	24658	22386	zh-en	21597	13768
en-tl	34980	31463			

Table 6: Number of bilingual word pairs (lines) and number of unique source language words (types) for each language pair in the MUSE word translation data. The ratio between lines and types can be seen as a measure of the average fertility (number of translations) of the source language words.

We split the data into three parts corresponding to frequency ranges of the source language words: The high frequency range provides bilingual word pairs where the frequency is among the 5000 most frequent words in the MUSE data. The mid frequency range consists of words ranking between 5001 and 20000, and the low frequency range belongs to ranks 20001 to 50000. However, for languages where the MUSE data comprises less than 50000 unique source language words (see Table 6), we had to reduce these thresholds. For en-ru and ru-en the thresholds were set to 5000, 20000 and 40000. For en-zh they are at 5000, 15000 and 30000, and for zh-en they are at 4500, 9000 and 13500.

From these ranges we extracted (pseudo) random samples which we call bins. Each bin comprises 2000 unique source language words together with all their translations. Like in the original MUSE data, also in the bins the source language words are ordered according to frequency (most frequent first). All three sets (per language pair) taken together, this gives 6000 unique source language words together with their translations, whereby, as shown in Table 5, each possible translation is listed in a separate line along with the source language word.

Given large datasets and an ambitious shared task schedule, we did not have the time to manually correct the data files. However, although the MUSE dictionaries were apparently generated automatically, they seem mostly of reasonably good quality, with only few errors. An exception is the low frequency range of English-Chinese where almost all source language words are translated by identical target language words which is not very useful. We encouraged the participants of the shared task to report to us such errors so that, as a positive side effect of the shared task, information for the improvement of the datasets was collected. For details, see the system description papers of the shared task participants in this volume.

For testing the systems, lists of source language test words were provided which, based on word frequency, were likewise split into three sets of 2000 unique words.

We had informed the participants that if their algorithms required a seed lexicon, they should use an arbitrary part of the training data for this purpose. Our hope was that with its 6000 source language words and even more translation pairs, the training set was large enough to provide for the participants' needs. If not, participants were referred to the open track of the shared task.

4. Open Track

In this track, participants were free to work on other language pairs, use their own data and, if desired, use their own evaluation procedures. They were encouraged to describe in their papers the reasons and motivation for deviating from the procedures of the closed track and, if possible, to provide access to their data. We also indicated that we might be able to give support for other language pairs by providing cleaned-up Wikipedia corpora and datasets of word translations extracted from MUSE.

Note that the limited choice of language pairs in the closed track was deliberate in order not to scatter participation over too many languages with the consequence of making comparisons between systems difficult. But in principle we

were prepared to offer support for all language pairs covered by the MUSE dictionaries.

As this appears to be the first shared task on the topic of dictionary induction from comparable corpora, we could not draw on previous experiences. Due to this pilot character, in Track 1 we were trying to keep things as clear and unsophisticated as possible. But in Track 2 we encouraged participants to challenge this simplicity, to freely experiment and to come up with new ideas in the hope that the resulting insights will promote future progress in the field.

5. Participants and Systems

Despite the ambitious schedule of the shared task, four teams managed to submit their results in time. These teams and the tracks and language pairs they worked on are listed in Table 7. As cited in the table, the first three teams have system description papers in this volume, which is why we only briefly describe their approaches here.

Short name	Long name / paper	Track and language pairs
CEN	Amrita School of Engineering, Center for Computational Engineering and Networking (CEN) (Sanjanasri et al., 2020)	*closed track:* de-en *open track:* ta-en
LMU	LMU Munich, Center for Information and Language Processing (Severini et al., 2020)	*closed track:* de-en, en-de, en-ru, ru-en *open track:* de-en, en-de, en-ru, ru-en
LS2N	Université de Nantes, TALN/LS2N (Laville et al., 2020)	*closed track:* de-en, en-de, de-fr, fr-de, en-es, es-en, en-fr, fr-en
SW	Sida Wang[8]	*closed track:* en-zh, zh-en

Table 7: Participating teams and their tracks and language pairs.

The LMU team relies on bilingual word embeddings which they claim to be effective in low resource settings. However, as they typically do not perform well on low frequency words, the embeddings are supplemented utilizing word surface similarity such as orthography and transliteration information.

The LS2N team combines a word embedding approach with a concatenation approach based on Tomas Mikolov's well known Word2vec[9] system together with a cognates matching approach based on string similarity.

The CEN team puts an emphasis on the transfer learning of semantics based on cross-lingual embeddings. For this purpose they experiment with different approaches, such as Word2Vec, Multilayer Perceptrons and Convolutional Neural Networks.

Sida Wang described his system as follows:[10]

"1) The system does not use the training data for training, instead it uses identical mappings as initialization and uses the training set as a validation set

2) An iterative procedure is used to figure out as much of the vocabulary as possible, independent of what is needed in the output (i.e. independent of the test set)
2a) I used the supervised rotation method where nearest neighbors (corrected with CSLS) are predicted as translations
2b) The iterative procedure adds (s,t) if t € top_k(s) and t € top k(t) where a k of 2 did the best on the validation set

3) My implementation is based on vecmap (https://github.com/artetxem/vecmap) but I only used a supervised procedure and a different iterative procedure as described above"

6. Evaluation

For evaluation, participants of the closed track (for the open track this was optional) were asked to provide their results on the test data sets for the test words in each of the three frequency ranges. Hereby it was expected that for each source language word all its major translations were provided (whereby the definition of "major" was supposed to be inferred from the training data). These translations were compared to the translations as found in the (internal) gold standard data which is structurally similar to the training data as it was randomly sampled from the same MUSE data in the same three frequency ranges. Only identical strings were considered correct, and the performance of a system was determined by computing precision (P), recall (R), and F1-score, the latter being the official score for system ranking. All data sets are in UTF-8 encoding.

More precisely: the input to the system is a list of source language words, one per line. A system was supposed to return, for each input word one or more candidate translations, in the form of tab-separated word pairs, each on its own line. For instance, in the English-French case, given the gold standard, test word list, and system output as shown in Table 8, the system would get credited for two true positives, one false positive, and two false negatives, hence

$$P = 2 / 3 = 0.67$$

$$R = 2 / 4 = 0.50$$

$$F1 = 2 (P * R) / (P + R) = 0.57$$

[8] http://www.sidaw.xyz/, https://www.linkedin.com/in/sidaw

[9] https://en.wikipedia.org/wiki/Word2vec

[10] E-mail to shared task organizers (May 2, 2020).

Table 9 shows some pseudo-code for computing these scores in a very simple and efficient way. The implementation can be conducted using standard UNIX commands such as *sort* and *wc*.

gold standard	
bed	lit
bed	plumard
doctor	médecin
doctor	docteur

test set
bed
doctor

system output	
bed	lit
bed	futon
doctor	docteur

Table 8: Sample gold standard, test word list and system output for the English-French case.

Inputs:
File with system output
File with gold standard data

Assumptions:
Tab-separated word pairs in both files (as in Table 1)
Only unique lines in both files (no repetitions)

Procedure:
A = number of lines in file with system output
B = number of lines in file with gold standard data
C = A + B
Merge both input files
Conduct unique sort of the lines in the merged file
D = number of lines in uniquely sorted file
NoMatches = C – D
R = NoMatches / B
P = NoMatches / A
F1 = 2 * (P * R) / (P + R)

Table 9: Pseudo code for computing recall, precision and F1-score.

7. Results

7.1 Overall results (without considering frequency bins)

Table 10 show the participating teams' results for the closed track. These are overall results not considering the frequency bins, i.e. when the data from the three frequency bins are merged for the gold standard data and also for the system output data. Table 11 shows analogous data for the open track. No evaluation was conducted for CEN's ta-en (Tamil-English) language pair as we had not provided a test set for this.

Overall results closed track				
Lang.	Team	R	P	F1
de-en	CEN	15.3	5.2	7.7
	LMU	48.7	61.6	54.4
	LS2N	57.5	66.2	61.5
en-de	LMU	40.2	59.8	48.1
	LS2N	54.3	54.8	54.5
en-ru	LMU	33.9	37.8	35.8
	LS2N[11]	32.6 *37.8*	38.7 *30.7*	35.4 *33.9*
ru-en	LMU	43.9	56.7	49.5
	LS2N	35.5	56.7	43.7
de-fr	LS2N	76.8	76.7	76.8
fr-de	LS2N	78.3	64.9	71.0
en-es	LS2N	63.8	61.4	62.6
es-en	LS2N	67.5	75.1	71.1
en-fr	LS2N	61.2	69.7	65.1
fr-en	LS2N	46.0	64.6	53.7
en-zh	SW	45.3	54.6	49.5
zh-en	SW	33.6	40.9	36.9

Table 10: Overall results for the closed track.

Overall results open track				
Lang.	Team	R	P	F1
de-en	LMU	50.6	63.8	56.4
en-de	LMU	41.1	61.1	49.2
en-ru	LMU	39.3	43.8	41.4
ru-en	LMU	50.7	65.4	57.1

Table 11: Overall results for the open track.

7.2 Results when considering frequency bins

Tables 12 to 15 show the teams' results when the high/mid/low frequency bins are distinguished. Again, no evaluation was conducted for CEN's ta-en (Tamil-English) language pair. Given the difficulty of the task where the teams not only had to rank candidates but also had to precisely decide which ones to keep and which ones to discard, we found the best results surprisingly good.

Concerning the frequencies of the source language words, often the results get better with lower frequencies, showing that the methods are quite good in dealing with sparse data. Only the low frequency words of the language pair zh-en, with an astonishing F1-score of 0.852, benefits from an idiosyncrasy of the MUSE data: Here almost all items consist of identical strings on the source and target language sides, which is particularly beneficial for the approach used by Sida Wang (see section 5).

[11] Normal font: Results based on overall file (no distinction of frequency bins) as provided by team. Italics: Results from merged high/mid/low-frequency bins. Bins provided by team.

Closed track by frequency										
Lang.	Team	high freq.			mid freq.			low freq.		
		R	P	F1	R	P	F1	R	P	F1
de-en	CEN	9.0	4.0	5.5	15.0	4.9	7.4	27.0	6.6	10.6
	LMU	44.7	49.1	46.8	43.4	70.9	53.8	62.8	77.1	69.2
	LS2N	48.1	63.7	54.8	59.0	63.0	60.9	72.2	73.3	72.8
en-de	LMU	35.1	51.4	41.7	35.0	65.3	45.6	61.4	71.2	66.0
	LS2N	49.0	51.6	50.3	53.7	52.6	53.2	68.6	65.2	66.9
en-ru	LMU	38.0	41.0	39.4	30.7	39.1	34.4	29.5	30.3	29.9
	LS2N	47.7	36.5	41.3	34.3	25.7	29.4	21.4	22.5	22.0
ru-en	LMU	45.3	67.6	54.2	45.5	59.4	51.5	39.9	43.1	41.4
	LS2N	49.3	59.0	53.7	38.3	56.0	45.5	13.2	48.8	20.8

Table 12: Comparison of results by frequency for the closed track.

Closed track by frequency LS2N									
Lang.	high freq.			mid freq.			low freq.		
	R	P	F1	R	P	F1	R	P	F1
de-fr	73.0	66.8	69.8	78.8	76.9	77.8	78.9	89.5	83.8
fr-de	73.9	50.2	59.8	79.1	67.0	72.6	82.0	85.9	83.9
en-es	57.6	61.7	59.6	63.3	56.8	59.9	77.8	67.1	72.1
es-en	61.9	74.9	67.8	66.4	72.8	69.4	77.2	78.0	77.6
en-fr	55.2	66.2	60.2	59.9	67.6	63.5	74.4	78.5	76.4
fr-en	54.6	65.6	59.6	49.1	64.3	55.7	29.4	62.0	39.8

Table 13: Results by frequency for the closed track for language pairs where only LS2N participated.

Closed track by frequency SW									
Lang.	high freq.			mid freq.			low freq.		
	R	P	F1	R	P	F1	R	P	F1
en-zh	39.1	40.9	40.0	27.0	41.5	32.7	78.1	93.8	85.2
zh-en	40.1	50.1	44.5	32.9	47.3	38.8	25.6	25.6	25.6

Table 14: Results by frequency for the closed track for language pairs where only SW participated.

Open track by frequency LMU									
Lang.	high freq.			mid freq.			low freq.		
	R	P	F1	R	P	F1	R	P	F1
de-en	44.6	49.0	46.6	46.7	76.2	57.9	66.4	80.8	72.9
en-de	35.4	51.8	42.0	36.8	68.6	47.9	62.5	72.3	67.1
en-ru	36.8	39.7	38.2	36.1	46.0	40.4	48.5	49.9	49.2
ru-en	46.9	70.0	56.2	50.3	65.5	56.9	56.3	60.7	58.4

Table 15: Results by frequency for the open track for language pairs where only LMU participated.

8. Conclusions and Outlook

The fourth BUCC shared task addressed the extraction of bilingual dictionaries from comparable corpora. This is a difficult task as, in contrast to parallel corpora, in this case it is not clear how to bridge the gap between languages. Nevertheless, the best participating systems achieved consistently good results for a number of language pairs involving languages from related as well as from very distant languages.

Of course, the provided datasets were not perfect: They were based on the automatically created MUSE dictionaries and, due to their considerable sizes, not manually checked. For each of 28 language pairs they comprised 12000 unique source language words (6000 for the training sets and another 6000 for the test sets) with somewhat more translations.

Challenges of interest for future shared tasks on bilingual lexicon induction from comparable corpora include:

1) Finding mappings across the full set of inflected forms of two languages. For example, *adequate* in English maps to four cognate forms in Spanish: *adecuado, adecuada, adecuados, adecuadas*, corresponding to the choices of singular vs. plural and feminine vs. masculine, because the English adjectives do not inflect for number and gender. The gold standard we used in the current shared task did not necessarily include the full range of forms.
2) Another issue concerns the representation of word senses in the test set. Since the gold standard translations were extracted from parallel corpora, as word selection in the target language is biased by the words in the source language, their set is likely to be different from what is available in general comparable corpora, such as the WaCky corpora and Wikipedia. For example, translations of *strong voice* extracted from the Europarl corpus primarily include references to expressions of opinions rather than assessments of the vocal cord. Translations also exhibit a cline from clear homonymy for words like *bank* to clear polysemy for words like *heavy* in which the same sense can be translated slightly differently depending on the context *heavy luggage, heavy blow, heavy rain*. More research is needed into what is the range of polysemous translations in the available test datasets.
3) In preparing data for this shared task we used information about the frequencies of words, as highly frequent words exhibit different translation properties from low frequent words. However, the test lexicon contains other sources of variation, which are worth a separate investigation, such as common names, borrowings or proper names. For example, borrowed proper names have sometimes trivial translations, e.g. *Kazimierz* maps to itself in the English to French evaluation set.
4) A particularly relevant topic is multiword expressions which are omnipresent in specialized language. We did not address them at all here, but this should certainly be a fruitful direction of research in the future.

9. Acknowledgements

We would like to thank Alexis Conneau from Facebook AI Research for allowing us to use extracts of the MUSE word translation data for the shared task and the Web-as-a-corpus initiative for providing the WaCky-corpora.

This work was partially funded by the Marie Curie Career Integration Grant MULTILEX within the 7th European Community Framework Programme and by the Marie Curie Individual Fellowship SEBAMAT within the European Commission's Horizon 2020 Framework Programme.

9. Bibliographical References

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano; Zanchetta, Eros (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43 (3): 209–22.

Bojanowski, Piotr; Grave, Edouard; Joulin, Armand; Mikolov, Tomas (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Vol. 5, 135 –146.

Conneau, Alexis; Lample, Guillaume; Ranzato, Marc'Aurelio; Denoyer, Ludovic; Jégou, Hervé (2017). Word translation without parallel data. *arXiv preprint arXiv: 1710.04087* (published at ICLR 2018).

Irvine, Ann; Callison-Burch, Chris (2017). A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics* 43 (2), 273–310.

Jakubina, Laurent; Langlais, Philippe (2016). A comparison of methods for identifying the translation of words in a comparable corpus: recipes and limits. *Computación y Sistemes* 20 (3), 449–458.

Laville, Martin; Hazem, Amir; Morin, Emmanuel (2020). TALN/LS2N participation at the BUCC shared task: bilingual dictionary induction from comparable corpora. *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*.

Martin, Joel; Mihalcea, Rada; Pedersen, Ted (2005). Word alignment for languages with scarce resources. *Proceedings of the ACL Workshop on Building and Using Parallel Texts*.

Mihalcea, Rada; Pedersen, Ted (2003). An evaluation exercise for word alignment. *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*.

Rapp, Reinhard; Sharoff, Serge; Zweigenbaum, Pierre (2016). Recent advances in machine translation using comparable corpora. *Journal of Natural Language Engineering* 22 (4), 501–516.

Sanjanasri, JP; Menon, Vijay Krishna; Soman, KP (2020). BUCC 2020: bilingual dictionary induction using cross-lingual embedding. *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*.

Severini, Silvia; Hangya, Viktor; Fraser, Alexander; Schütze, Hinrich (2020). LMU bilingual dictionary induction system with word surface similarity scores for BUCC 2020. *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*.

Sharoff, Serge; Goldhahn, Dirk; Quasthoff, Uwe (2017). *Frequency Dictionary: Russian*. Leipziger Universitätsverlag. http://corpus.leeds.ac.uk/serge/publications/2017-russian-frq-leipzig.pdf

Sharoff, Serge; Rapp, Reinhard; Zweigenbaum, Pierre (2013). Overviewing important aspects of the last twenty years of research in comparable corpora. In: Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, Pascale Fung (eds.): Building and Using Comparable Corpora. Heidelberg: Springer, 1–18.

Sharoff, Serge; Rapp, Reinhard; Zweigenbaum, Pierre (2013). Overviewing important aspects of the last twenty years of research in comparable corpora. In: Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, Pascale Fung (eds.): *Building and Using Comparable Corpora*. Heidelberg: Springer, 1–18.

Sharoff, Serge; Zweigenbaum, Pierre; Rapp, Reinhard (2015). BUCC shared task: cross-language document similarity. *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, Beijing, China. ACL Anthology, 74–78, http://www.aclweb.org/anthology/W15-3411.pdf