Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022 , pages 67–76
Marseille, June 25, 2022

Overview of the 2022 BUCC Shared Task: Bilingual Term Alignment in Comparable Specialized Corpora

Omar Adjali[1], Emmanuel Morin[2], Serge Sharoff[3], Reinhard Rapp[4], Pierre Zweigenbaum[1]
[1]Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France
[2]Nantes Université, CNRS, Laboratoire des Sciences du Numérique de Nantes, France
[3]University of Leeds, Leeds, UK
[4]University of Mainz, Germany
omar.adjali@universite-paris-saclay.fr, pz@lisn.fr,
emmanuel.morin@univ-nantes.fr, s.sharoff@leeds.ac.uk, reinhardrapp@gmx.de

Abstract

The BUCC 2022 shared task addressed bilingual terminology alignment in comparable corpora. Many research groups are working on this problem using a wide variety of approaches. However, as there is 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. The shared task aimed at solving these problems by organizing a fair comparison of systems. This was accomplished by providing a precise definition of the task and its evaluation, and corpora and non-trivial evaluation datasets for the English-French language pair. Six runs were submitted by two teams. The obtained results are satisfactory with a top average precision of 0.28, but show that the task is not solved yet.

1. Introduction

The abundance of web data makes comparable corpora texts readily available in many language pairs including under-resourced languages and for specialized domains. This encouraged the natural language processing (NLP) community to investigate how they could benefit the development of machine translation models and related tasks, thus alleviating the scarcity of parallel resources. Unlike parallel corpora where texts are in a strong translation relation (strongly comparable), texts in comparable corpora are either weakly comparable by covering the same topics and domains, or can be totally unrelated (Sharoff et al., 2013). Although these corpora exhibit only weak parallelism, the inherent cross-lingual information in comparable corpora is sufficient to enhance statistical machine translation (Abdul-Rauf and Schwenk, 2009; Irvine and Callison-Burch, 2013; Rapp et al., 2016), bilingual terminology extraction (Fišer and Ljubešić, 2011; Ljubešic et al., 2012; Aker et al., 2013; Hazem and Morin, 2016), as well as development of multilingual pre-trained language models (Devlin et al., 2019; Conneau et al., 2020).

The first shared task in the BUCC workshops series on Building and Using Comparable Corpora addressed the cross-lingual detection of comparable documents (Sharoff et al., 2015). This task paved the way for several others which compensated for the lack of shared tasks on the topic of comparable corpora. The second and third BUCC shared task (Zweigenbaum et al., 2016; Zweigenbaum et al., 2017; Zweigenbaum et al., 2018) proposed to tackle the extraction of parallel sentences from comparable corpora. The fourth BUCC shared task (Rapp et al., 2020) addressed bilingual dictionary induction from comparable corpora.

The 2022 BUCC workshop follows theses endeavours on building and using comparable corpora, and proposes a shared task on bilingual terminology alignment in comparable specialized corpora. Here, we seek to evaluate methods that detect pairs of terms that are translations of each other in two comparable corpora, with an emphasis on multi-word terms in specialized domains. While a variety of approaches have been proposed to solve the bilingual terminology alignment problem, it becomes difficult to assess and compare them as there is no standard way to measure the performance of the systems. The published results are therefore not comparable and the pros and cons of the various approaches are not clear. An important cause is the difference in experimental settings, such as the language pair, the specificity of the comparable corpora, the size of the dataset, the occurrence of multi-word terms, and what is considered a valid translation.

The present shared task aimed at solving these problems by organizing a fair comparison of systems. This was accomplished by providing corpora and evaluation datasets for the English-French language pair. Furthermore, the importance of dealing with multi-word terms (MWTs) in Natural Language Processing applications has been recognized for a long time. In particular, multi-word terms pose serious challenges for machine translation systems because of their syntactic and semantic properties. They also tend to be more frequent in domain-specific text: some studies established that multi-word terms represent the largest proportion of lexical units in a domain-specific lexicon (Constant et al., 2017), hence the need to handle them in tasks with specialized-domain corpora.

In this paper, we report the present task as a companion to the BUCC 2022 workshop. We explain how we defined the task (Section 2), how we prepared the data and built the datasets (Section 3), and how we evaluated the task (Section 4). We then present the participants' systems and their evaluation results (Section 5).

2. Task Design

Bilingual terminology extraction from comparable corpora in two languages L_1 and L_2 can be broken down into three subtasks:

1. Collecting comparable corpora C_1 and C_2 for languages L_1 and L_2.
2. Extracting monolingual terms from corpus C_1 (resp. C_2), leading to monolingual term set D_1 (resp. D_2).
3. **Aligning bilingual terms**, i.e., finding term pairs $(t_1, t_2) \in D_1 \times D_2$ such that t_1 and t_2 are translations of each other.

We aimed to remove the variance due to corpus collection and term extraction methods and focus on the multilingual step in the process: bilingual term alignment. In this purpose, we performed the corpus collection and monolingual term extraction steps ourselves, and provided predefined corpora and term lists to task participants.

Furthermore, to evaluate the list of term pairs found by a system, we provided a gold standard bilingual dictionary $D_{1,2} \in D_1 \times D_2$. We detail in Section 4 the chosen evaluation rationale and method. It aims to take into account the confidence a given alignment method has in the term pairs it proposes, hence expects a system to produce a term list ranked in decreasing order of confidence.

This led to the following shared task definition. Given a pair of comparable corpora (C_1, C_2) in two different languages (L_1, L_2), and a set of terms D_1 found in C_1 and a set of terms D_2 found in C_2, the objective is to produce a list of term pairs $(t_1, t_2) \in D_1 \times D_2$ that are translations of each other, in descending order of confidence. Note that D_1 and D_2 may have different sizes, that not every term in D_1 may have a translation in D_2, that some terms in D_1 might have multiple translations, and conversely.

Additionally, for practical reasons, we limited the length of a submitted term pair list to a ceiling of 10 times the average length of D_1 and D_2. This can be seen as meaning that, on average, a system could submit up to 10 alignment hypotheses for each term in D_1 or in D_2.

3. Data preparation

As explained above, the BUCC 2022 shared task dataset required the preparation of the following data:

- A pair of comparable corpora (C_1, C_2) in languages (L_1, L_2).
- A set of terms $D_1 = \{t_1^i$ occurring in $C_1\}$ and a set of terms $D_2 = \{t_2^j$ occurring in $C_2\}$.
- A gold standard dictionary $D_{1,2}$ in the form of a set of pairs of terms (t_1^i, t_2^j) in $D_1 \times D_2$ that are translations of each other.

A straightforward method to build a gold standard dictionary to measure the performance of bilingual alignment systems would be to manually extract bilingual terms from the given comparable corpora. This would require tremendous manual annotation efforts. To sidestep this issue, we built the gold standard dictionary $D_{1,2}$ by performing bilingual term extraction from a parallel corpus. Such an approach (Arcan et al., 2014; Yang et al., 2016; Krstev et al., 2018; Šandrih et al., 2020) leverages sentence-level/document-level alignment information that helps retrieve cross-lingual signals between pairs of bilingual terms. Thus, we first extracted monolingual terms from a large parallel corpus of aligned sentence pairs that are translations of each other (El-Kishky et al., 2020). This yielded two preliminary lists of terms that we used as resources to build the gold standard $D_{1,2}$. We summarize the process below; more detail can be found in (Adjali et al., 2022).

3.1. Parallel Corpora

CCAligned is a large sentence-aligned dataset built from sixty-eight snapshots of the Common Crawl corpus (El-Kishky et al., 2020). In this dataset, pairs of web documents that are translations of each other have been identified in 8,144 language pairs, of which 137 pairs include English. El-Kishky et al. (2020) identified each document's language using a text classifier (fastText), and identified pairs of cross-lingual documents using a high-precision, low-recall heuristic to assess whether two URLs represent web pages that are translations of each other. To assess their dataset construction approach, they ran a human evaluation on a diverse sample of positively-labeled documents across six language pairs. We used the English-French parallel corpus that contains 15,502,845 aligned sentence pairs.

3.2. Building Specialized Comparable Corpora from Parallel Corpora

We extracted comparable corpora from the CCAligned parallel corpus to provide resources for training and evaluating term alignment systems. Specifically, we de-parallelized pairs of aligned sentences by discarding one of the two sentences in each sentence pair. Given an aligned sentence pair (L_1,L_2), the L_1 sentence was discarded with probability p or the L_2 sentence with probability 1-p. Moreover, the large size of CCaligned allowed us to sample thematic comparable sub-corpora. We first investigated topic modelling techniques, however they failed to output satisfying results. We believe that topic models perform

well on document-level corpora, whereas our input was sentence-level corpora. We therefore resorted to information retrieval methods, and used seed lexicons found in external resources as keyword queries to select sentences and build specialized comparable sub-corpora.
We specifically used the Medical Subject Headings (MeSH) terminology (27,456 entries) as an input seed. Following the procedure described above for generating non-parallel corpora, we generated the corresponding comparable corpora with sentences containing MeSH terms.
Table 1 shows example selected sentences that illustrate the occurrence in non-parallel sentences of terms and their translations. The bold text shows the aligned terms from the gold standard dictionary.

3.3. Monolingual Term Extraction

We used *TermSuite* (Rocheteau and Daille, 2011; Cram and Daille, 2016) for automatic term extraction (ATE). TermSuite is a multilingual terminology extractor that identifies term candidates using linguistic (morphosyntactic regular expression) and statistical information. Using syntactic and morphological variant patterns, it also performs term variant recognition which enriches the output of extracted terms. We performed automatic term extraction on the CCAligned English-French sentence pairs and collected all terms. This provided lists of monolingual terms that represent our source and target term candidates. We discarded terms containing proper names. Table 2 shows the number of terms and multiword terms (MWTs) after monolingual ATE.

3.4. Bilingual Term Alignment

We used two complementary methods to find translation correspondences between source terms and target terms. First, we employed an embedding-based approach which consists in mapping word representations of each language, learnt separately from monolingual corpora, into a common vector space, and then using these embeddings to align a source term with the closest target term in the embedding space. In parallel, we performed statistical machine translation on the source term list extracted during monolingual ATE. Both methods produced term alignments that supported a final manual annotation step. We detail these two methods and the manual step below.

Embedding-based Alignment

We learned vector representations of the source and target terms using the Compositional Approach with Word Embedding Projection (CMWEP) approach proposed in (Liu et al., 2018b). They extended the linear matrix transformations approach of (Zhang et al., 2016; Liu et al., 2018a; Artetxe et al., 2018) which learns a mapping matrix that projects source word embeddings to the target embedding space. In a nutshell, we used the English and French 300-D fastText word vectors trained on Common Crawl and Wikipedia (Bojanowski et al., 2016). Using these and following the linear transformation approach (Artetxe et al., 2016), we learned the mapping matrix that projects source word vector representations to the target vector space. Then, using a seed bilingual dictionary (Conneau et al., 2017), source terms in D_1 were represented using the embedding vectors of their translation found in the dictionary, while out-of-vocabulary source term representations were computed using the mapping matrix. Note that by averaging the vector representations of their component words, the CMWEP approach assumes the compositionality of multi-word terms. Finally, we used cosine similarity as an alignment score between the vector representations of each term in the source list D_1 and the vector representations of their corresponding target candidates in D_2. The candidate translations were then ranked according to their scores.

Statistical Machine Translation

We used the Moses (Koehn et al., 2007) Statistical Machine Translation (SMT) system to translate the source terms. We trained the translation model using the CCAligned English-French parallel corpus, and its language model using the French sentences. For each source term in D_1, we generated the 10 best translations.

Manual Annotation

The embedding-based alignment of terms in D_1 and D_2 yielded a preliminary large translation dictionary from which we randomly selected a subset of term pairs that occurred in the extracted specialized comparable corpora. To avoid unnecessary manual verification, we automatically discarded all term pairs with embedding alignment scores (cosine similarity) below a threshold set to 0.6, since we empirically observed that most term pairs with lower similarity scores were misaligned. For the remaining term pairs, we checked for each source term whether the embedding-based aligned target term was a correct translation. Additionally, we systematically examined the 10 best translations given by the SMT system, looking for alternative correct translations to add to the gold dictionary. A source term can thus have multiple translations. Instances include term pairs (*design possibilities, possibilités de conception*), (*design possibilities, possibilités de design*) and term pairs (*anti-cancer, anti-cancer*), (*anti-cancer, anticancéreux*).
To make the alignment task more challenging, we considered the following aspects:

- We added to the lists of source and target terms mis-aligned term pairs with high embedding-based alignment scores. Indeed, embedding-based alignment approaches may fail when target candidate representations are close in the embedding space.
- Conversely, when term pairs were correctly aligned, we examined the closest target term candidates in the embedding space and included in

English comparable corpus samples	French comparable corpus samples
health care professionals are eligible to log-on to view their patients' records. That's a 109% increase since 2016.	Etude de marché réalisée par The Research Partnership, Ltd en février 2018 en Europe et aux US sur un échantillon de 75 **professionnels de santé.**
A range of quality stethoscopes to meet the needs of all **health care professionals** and paramedical staff.	Afin de célébrer ses 20 ans d'engagement au service de votre santé, le Laboratoire LESCUYER a réalisé, le 7 juin 2014, son 1er Symposium de Micronutrition dédié aux **professionnels de santé**.
Our company started its activity in year 2002, being a young company, but with a team in constant development and learning, and opened to the new systems of treatment and valuation of **industrial wastes**.	Le traitement des **déchets industriels** est très coûteux pour les entreprises, en plus d'être un problème environnemental important en raison de leur volume et caractéristiques physico-chimiques.
Effective technology to treat liquid **industrial waste**, which can create a zero-dumping treatment system. It enables waste to be minimised by concentration, which reduces the cost of managing it.	Chez Yamasa Corporation, nous séparons les **déchets industriels** et nous nous efforçons de les réduire et les recycler le plus possible.
For the first time in ten years, the European Commission is considering increasing action on global **deforestation**, going beyond the problem of illegal logging.	Compte tenu de l'évolution rapide de la **déforestation** dans les espèces aire de répartition restreinte, et sa forte dépendance sur les habitats forestiers soupçonné sa population est en déclin rapide.

Table 1: English-French comparable corpora: example sentences

Lang	#Terms	#MWTs
En	130,681	48,889
Fr	286,581	59,529

Table 2: Monolingual automatic term extraction statistics

Nbr of	Train	Test
gold pairs	2,519	1,970
English sentences	1,148,695	221,838
French sentences	1,161,269	723,515
English terms	3,132	1,270
French terms	2,984	9,712

Table 3: Dataset and corpus statistics for each split

the list of target terms those that we considered as potentially confusing terms. As an example, given the English source term *infection*, we added term variants such as *infections bactériennes* and *infections respiratoires*. Note that we applied the same principle to the 10 best translation terms provided by the SMT system as long as the translated terms occurred in the comparable corpora.

- We took into account grammatical properties in the alignment: for example source terms in the plural form were aligned with target terms also in the plural form.

We altogether built two gold standard dictionaries for the training and test sets, which included respectively 2,519 and 1,970 term pairs (see Table 3).

4. Evaluation

4.1. Evaluation Metric: Average Precision

We modeled the bilingual term alignment task as an Information Retrieval (IR) task with one query that returns a ranked list of term pairs, as specified in the task definition (see Section 2). That task consists in retrieving all relevant term pairs (t_1, t_2) (each pair is a document in the IR evaluation setting) from the cross-product $D_1 \times D_2$ (virtual pool of documents), presenting them in descending order of confidence.[1]

The official evaluation metric for the BUCC 2022 Shared Task was the Average Precision of the predicted bilingual term pair list, where the relevance of a term pair is determined by its presence in the (hidden) gold standard dictionary $D_{1,2}$.

Average Precision is the area under the recall $\times$ precision curve. As shown in Eq. 1, it is computed as the average over all m relevant term pairs (t_i, t_j) (i.e., all term pairs in the gold standard) of the precision scores $P(R_k)$; R_k is the set of n_k term pairs retrieved by the system when it has retrieved k relevant term pairs. $P(R_k)$ is the precision of that set of term pairs, as defined in Eq. 2, where $D_{1,2}$ is the gold standard dictio-

[1]Note that with this definition of the task, term pairs must belong to $D_1 \times D_2$. When a system produced some term pairs outside $D_1 \times D_2$, we considered that this was due to an incorrect understanding of the task specification and did not penalize it: we only filtered out these 'out-of-vocabulary' term pairs before evaluating its output.

nary.

$$AP = \frac{1}{m} \sum_{k=1}^{m} P(R_k) \quad (1)$$

$$P(R) = \frac{|R \cap D_{1,2}|}{|R|} \quad (2)$$

Relevant term pairs that are not retrieved by the system receive a precision of zero, hence decrease Average Precision.

4.2. Variants Evaluation Scores and their Interpretation

4.2.1. MAP

Average Precision is defined for one query. Information Retrieval usually evaluates systems on multiple queries, and computes the mean over queries of their individual average precision scores, known as Mean Average Precision (MAP), as shown in Eq. 3:

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q) \quad (3)$$

The bilingual term alignment task could have been defined as one subtask per source term (viz. one subtask per target term), consisting of ranking target terms as candidate translations for that source term. Systems would have needed to return a ranked list of target terms for each source term, and MAP would have been a relevant evaluation score.

The main difference between the task defined in BUCC 2022, evaluated with Average Precision, and the variant discussed here, evaluated with MAP, is that by averaging over source terms, this variant would give the same weight to each source term, since it macro-averages over queries, i.e., source terms.

In contrast, the BUCC 2022 definition lets a system give priority to term pairs that involve any source or target term, in the order that it deems most appropriate given the confidence it has in these alignments. Average Precision gives a lower penalty to false positives that are ranked later by a system. If a system has a low confidence in a given alignment, it is better for it to push it towards the end of its list of retrieved term pairs, and to promote term pairs with a higher confidence. This is in line with a use case of collecting a large number of high-confidence bilingual term alignments from a given pair of comparable corpora.

4.3. Interpolation

Between two successive, but non-contiguous values of k in Eq. 1, precision will drop. The standard way to remove these drops is to turn to an interpolated precision $P_{interp}(R_k)$. Interpolated precision at a certain recall level R_k is defined as the highest precision found for any recall level $R_{k'} \geq R_k$, i.e., $k' \geq k$, as expressed by Eq. 4:

$$P_{interp}(R_k) = \max_{k' \geq k} P(R_{k'}) \quad (4)$$

Interpolated Average Precision is sometimes seen as too optimistic, therefore the official score in the present task was computed with uninterpolated Average Precision.

4.4. Computing Average Precision

We considered the following tools which provide functions to compute Average Precision or MAP, and eventually implemented it directly.

Scikit-Learn provides the `average_precision_score` function in its `sklearn.metrics` module. However, this function expects a system to rank (actually, score) every document in the collection. It is therefore not ideal for an Information Retrieval context, in which systems are free to retrieve and rank only a subset of the documents. In the case of the BUCC 2022 Shared Task too, systems were not expected to rank all possible term pairs in the cross-product of source and target terms; specifically, they were not expected either to include every gold term pair in their ranking, which is required by `average_precision_score`.

trec_eval computes the MAP score among a wealth of other scores. MAP is designed to score multiple queries, but if applied to one query, it effectively computes Average Precision. However, Information Retrieval often only evaluates the top N documents returned by a system for an individual query (e.g., $N = 1000$), and `trec_eval` was designed to set limits to that number. We were unsure whether we correctly directed it to raise that limit, so we also discarded `trec_eval` from our list.

Eventually, we directly implemented a function to compute uninterpolated and interpolated Average Precision. We also added functionality to plot the graph of Average Precision across a system's ranked list of results. The resulting code is available online.[2]

5. Systems and Results

5.1. Shared Task Systems

Two teams (see Table 4) submitted runs and proposed several approaches for bilingual term alignment. The

[2] `https://github.com/PierreZweigenbaum/bucc2022`

System	Affiliation
CUNI	Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (Požár et al., 2022)
IJS	Jožef Stefan Institute, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia (Repar et al., 2022)

Table 4: Shared Task Participants

run	AP	nSys	nGold	TP	FP	FN	P	R	F1
$CUNI_{\text{combined}}$	**0.2816**	6550	1970	945	5605	1025	0.1443	0.4797	0.2218
$CUNI_{\text{monoses}}$	0.1893	15477	1970	757	14720	1213	0.0489	0.3843	0.0868
$CUNI_{\text{muse}}$	0.1673	1570	1970	625	**945**	1345	**0.3981**	0.3173	**0.3531**
IJS_1	0.0054	14974	1970	366	14608	1604	0.0244	0.1858	0.0432
IJS_2	0.2674	54860	1970	**1576**	53284	**394**	0.0287	**0.8000**	0.0555
IJS_3	0.2685	54851	1970	1368	53483	602	0.0249	0.6944	0.0482

Table 5: Scores of the six submitted runs on the test set, in alphabetical order of run name: uninterpolated Average Precision, number of term pairs ranked by the system, number of gold standard term pairs, true positives, false positives, false negatives, precision, recall, F1-score. Bold shows the best result, underline the second best.

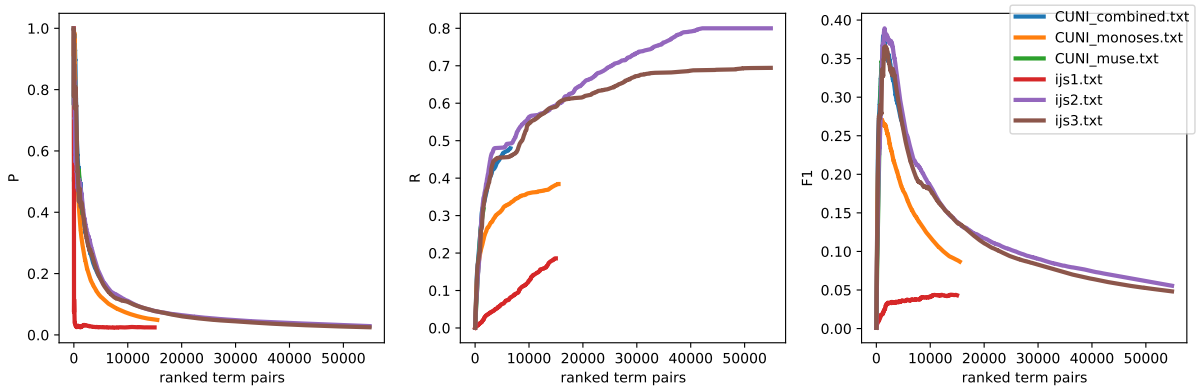

Figure 1: Precision, recall, and F1-score against the number of examined term pairs

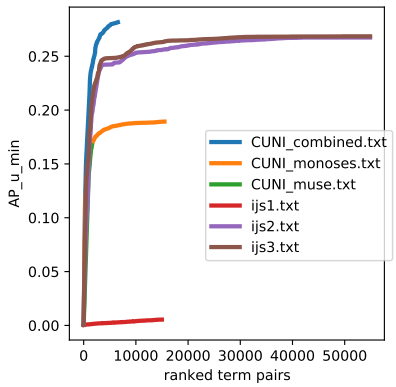

Figure 2: Uninterpolated average precision against the number of examined term pairs

CUNI team (Požár et al., 2022) submitted three approaches: a cross-lingual embedding-based approach using the MUSE tool (Conneau et al., 2017), a statistical phrase-based machine translation approach using the monoses pipeline (Artetxe et al., 2019) and contextualized embeddings from models such as multilingual BERT (Devlin et al., 2019). The IJS team (Repar et al., 2022) submitted three settings of an SVM-based machine learning approach that combines multiple features including sentence embeddings and dictionary-based features. Table 5 reports their results on the test set, after filtering out-of-vocabulary terms if necessary. We observed only tiny differences when computing uninterpolated and interpolated average precision for the participant system runs, so we only report the uninterpolated versions here.

5.2. Discussion

Overall, two of the IJS system settings (let us refer to them collectively as $IJS_{2,3}$) output considerably more suggestions, thus producing considerably more true positives, and a higher recall, but this comes at the expense of precision, which drives down both the AP and F1 scores. Conversely, two of the CUNI methods (MUSE and COMBINED) produce a much shorter list of term pairs which leads to a higher precision and F1-score.

Precision, recall, and F1-score are set measures that are computed on the whole set of returned term pairs, without taking their ranks into account. Conversely, average precision takes into account the ranks of the returned term pairs. It is thus better analyzed by plotting the evolution of precision, recall and F1-score (Fig. 1) and average precision (Fig. 2) as term pairs are examined in the order in which they are ranked by a system.

Fig. 2 shows that the $IJS_{2,3}$ methods, although returning a much longer list of term pairs than the CUNI_combined method, rank true positives less close to the top of the list. Though they have a much higher

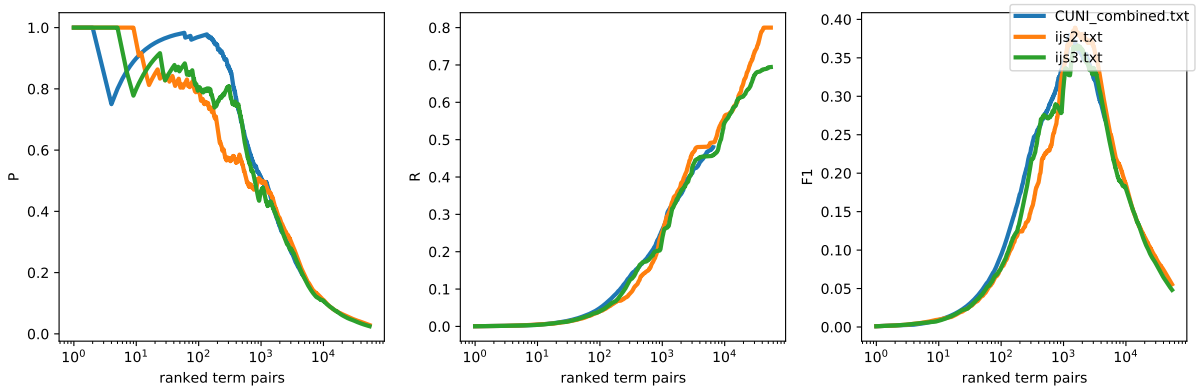

Figure 3: Precision, recall, and F1-score against the number of examined term pairs, with log X scale, for the systems with the top three average precisions: CUNI_combined and $IJS_{2,3}$

recall, their average precision is slightly lower than that of CUNI_combined. A more detailed inquiry of the top ranks thanks to a logarithmic X scale (Fig. 3) reveals that CUNI_combined generally has a better precision than $IJS_{2,3}$ in the top 1000 ranks. This confers it an initial advance in average precision which is not compensated by the larger, but later, number of true positives collected by $IJS_{2,3}$. This emphasizes the importance that average precision puts on obtaining high precisions early on and as long as possible: although these three systems (as well as CUNI_muse) reach a precision at 1000 of about 0.5, their paths to this score are quite different and result in a higher average precision at that point for CUNI_combined.

All in all, the average precision gap between $IJS_{2,3}$ and CUNI_combined is very small, and these two sets of systems illustrate two ways a certain level of average precision can be obtained: higher precision early on, and higher recall through a longer tail. The CUNI_combined method might now aim to increase its number of ranked term pairs, thus potentially continuing to collect true positives and expanding its average precision score. The $IJS_{2,3}$ method would need conversely to increase the precision of its higher-ranked term pairs.

We used the differential evaluation method of (Gianola et al., 2021) to estimate how difficult each gold term pair was to find by the participant systems. According to this method, an instance (here, a term pair) is considered all the more difficult as less systems are able to find it. With six evaluated systems, we partition the gold term pairs into seven bins: bin_n contains the term pairs that were found by exactly n systems, with $n \in \{0, 1, \ldots, 6\}$. Term pairs in bin_6 are considered the easiest because they were found by all systems, whereas term pairs in bin_0 are considered the hardest because no system was able to spot them. Table 6 shows the size of each bin. 10.7% of the gold terms were found by no system, 10.4% were only found by one system; only 4.6% were spotted by all six systems, and only 17.9% by at least five systems. This shows that the dataset did not prove an easy one for this set of systems.

bin	0	1	2	3	4	5	6	Total
size	211	205	431	383	369	281	90	1970
%	10.7	10.4	21.9	19.4	18.7	14.3	4.6	100

Table 6: Number of term pairs in each difficulty bin

Table 7 shows examples of term pairs in bin_0. These 'hard' term pairs illustrate several aspects of the task. The term *twin sister*$_{\text{en}}$ is singular, hence should be translated by a singular term. CUNI_combined found the plural *sœurs jumelles*$_{\text{fr}}$, but the gold standard only admitted the singular *sœur jumelle*$_{\text{fr}}$. All systems found the correct alignment *joystick*$_{\text{fr}}$ for *joystick*$_{\text{en}}$, but none of them identified the commonly used *manette*$_{\text{fr}}$ present in the gold standard. IJS_2 found *manche*$_{\text{fr}}$, which is close to it and is proposed by several on-line dictionaries: this term should be considered to extend the gold standard. No system found an approaching alignment for *ground-based*$_{\text{en}}$, maybe because the syntactic pattern of its gold alignment *au sol*$_{\text{fr}}$ (literally: *on the ground*$_{\text{en}}$) is less frequent among terms. None of the three gold alignments for *lust*$_{\text{en}}$ were identified by any system; IJS_1 found the approaching *passion*$_{\text{fr}}$ which is

en	fr	system
twin sister	sœur jumelle	sœurs jumelles
joystick	manette	manche
ground-based	au sol	—
lust	luxure	—
lust	convoitise	—
lust	désir	passion

Table 7: Example term pairs in bin_0, i.e., found by no system, and approaching alignments if available

also proposed by on-line dictionaries and thus a candidate to enter the gold standard dictionary. This difficulty might be caused by the fact that $lust_{\text{en}}$ encompasses multiple related senses that are translated as different French words. In that context, even though the gold dictionary contained three alignments, none of the tested methods was able to figure them out.
Conversely, Table 8 shows examples of term pairs found by all systems. They illustrate two properties of bin_6: all source and target terms in that bin are single-word terms, and in every pair, the source and target terms are morphologically similar. This typically makes the task easier for many methods.

en	fr
distancing	distanciation
confortable	confortable
algorithmic	algorithmique
blockchain	blockchain
leitmotiv	leitmotiv
diligent	diligente
dressing	dressing
reflexology	réflexologie
diffraction	diffraction

Table 8: Example term pairs in bin_6, i.e., found by all systems

These results were obtained on one test dataset; more datasets in more languages will be needed to obtain a broader view of the current performance of systems and of the remaining challenges.

6. Conclusion

The BUCC 2022 shared task addressed bilingual term alignment in comparable corpora. We described the semi-automatic method used to prepare the data comprised of comparable corpora and a bilingual terminology.
Six runs were submitted by two teams, with a top average precision of 0.28. Four systems obtained a precision@1000 of about 0.5. The system with the highest recall (0.80) obtained a precision of 0.03; the system with the highest precision (0.40) obtained a recall of 0.32. We studied examples of term pairs that proved easy or difficult for this set of systems. These figures and examples show that the task of bilingual term alignment in comparable corpora, as defined here and implemented through a dataset, is not yet solved.

Acknowledgements

This work has been funded by the French National Research Agency (ANR) under the ADDICTE project (ANR-17-CE23-0001).

7. Bibliographical References

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece, March. Association for Computational Linguistics.

Adjali, O., Morin, E., and Zweigenbaum, P. (2022). Building comparable corpora for assessing multiword term alignment. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France, May. European Language Resources Association.

Aker, A., Paramita, M. L., and Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–411.

Arcan, M., Turchi, M., Topelli, S., and Buitelaar, P. (2014). Enhancing statistical machine translation with bilingual terminology in a CAT environment. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 54–68, Vancouver, Canada, October 22-26. Association for Machine Translation in the Americas.

Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Artetxe, M., Labaka, G., and Agirre, E. (2018). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Artetxe, M., Labaka, G., and Agirre, E. (2019). An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., and Stoyanov, V. (2020). Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online, July. Association for Computational Linguistics.

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Cram, D. and Daille, B. (2016). Terminology extrac-

tion with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, pages 13–18.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (2020). A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

Fišer, D. and Ljubešić, N. (2011). Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 125–131.

Gianola, L., El Boukkouri, H., Grouin, C., Lavergne, T., Paroubek, P., and Zweigenbaum, P. (2021). Differential evaluation: a qualitative analysis of natural language processing system behavior based upon data resistance to processing. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 1–10, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Hazem, A. and Morin, E. (2016). Efficient data selection for bilingual terminology extraction from comparable corpora. In *26th International Conference on Computational Linguistics (COLING)*, pages 3401–3411.

Irvine, A. and Callison-Burch, C. (2013). Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the eighth workshop on statistical machine translation*, pages 262–270.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Krstev, C., Šandrih, B., Stanković, R., and Mladenović, M. (2018). Using English baits to catch Serbian multi-word terminology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Liu, J., Morin, E., and Saldarriaga, S. P. (2018a). Alignement de termes de longueur variable en corpus comparables spécialisés (alignment of variable length terms in specialized comparable corpora). In *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN*, pages 19–32.

Liu, J., Morin, E., and Saldarriaga, S. P. (2018b). Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2855–2866.

Ljubešic, N., Vintar, Š., and Fišer, D. (2012). Multiword term extraction from comparable corpora by combining contextual and constituent clues. In *The 5th Workshop on Building and Using Comparable Corpora*, page 143.

Požár, B., Tauchmanová, K., Neumannová, K., Kvapilíková, I., and Bojar, O. (2022). CUNI submission to the BUCC 2022 shared task on bilingual term alignment. In *Proceedings of the 15th Workshop on Building and Using Comparable Corpora*, pages 43–49, Marseille, France, June. European Language Resources Association.

Rapp, R., Sharoff, S., and Zweigenbaum, P. (2016). Recent advances in machine translation using comparable corpora. *Natural Language Engineering*, 22(4):501–516.

Rapp, R., Zweigenbaum, P., and Sharoff, S. (2020). Overview of the fourth BUCC shared task: Bilingual dictionary induction from comparable corpora. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 6–13, Marseille, France, May. European Language Resources Association.

Repar, A., Koloski, B., Ulčar, M., and Pollak, S. (2022). Fusion of linguistic, neural and sentence-transformer features for improved term alignment. In *Proceedings of the 15th Workshop on Building and Using Comparable Corpora*, pages 61–66, Marseille, France, June. European Language Resources Association.

Rocheteau, J. and Daille, B. (2011). TTC TermSuite - a UIMA application for multilingual terminology extraction from comparable corpora. In *Proceedings of the IJCNLP 2011 System Demonstrations*, pages 9–12, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Šandrih, B., Krstev, C., and Stanković, R. (2020). Two approaches to compilation of bilingual multiword terminology lists from lexical resources. *Natural Language Engineering*, 26(4):455–479.

Sharoff, S., Rapp, R., and Zweigenbaum, P. (2013). Overviewing important aspects of the last twenty years of research in comparable corpora. In *Building and using comparable corpora*, pages 1–17. Springer.

Sharoff, S., Zweigenbaum, P., and Rapp, R. (2015). BUCC shared task: Cross-language document similarity. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 74–78.

Yang, W., Yan, J., and Lepage, Y. (2016). Extraction of bilingual technical terms for Chinese-Japanese

patent translation. In *Proceedings of the NAACL Student Research Workshop*, pages 81–87.

Zhang, Y., Gaddy, D., Barzilay, R., and Jaakkola, T. (2016). Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, San Diego, California, June. Association for Computational Linguistics.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2016). Towards preparation of the second BUCC shared task: Detecting parallel sentences in comparable corpora. In *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora. European Language Resources Association (ELRA), Portoroz, Slovenia*, pages 38–43.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2017). Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2018). Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42.