1 Appendix

In this appendix, we provide additional analysis, results and examples.

1.1 Choice of languages and mined bitexts

We currently handle 90 languages. They were chosen to cover several language families, frequent high-resources languages as well as several low-resource languages for which only a very limited amounts of bitexts are publicly available. Mining all possible $90 \times (90-1)/2 \approx 4000$ pairs is computationally very challenging. In addition, one may question if mining several *"unusual"* language pairs would be useful, e.g. Icelandic-Urdu or Gallican-Malay. It is quite unlikely that there is interest in training a direct NMT systems for such language pairs.

Therefore, we took the following approach. We first organized all languages into twelve groups:

- Major Asian (4): Japanese, Korean, Vietnamese, Chinese;
- Germanic (12): Afrikaans, Danish, Dutch, German, English, Frisian, Icelandic, Luxembourgish, Norwegian, Swedish, Yiddish;
- Romance (10): Asturian, Catalan, French, Galician, Italian, Latin, Occitan, Portuguese, Romanian, Spanish.
- Slavic (12): Belarusian, Bosnian, Bulgarian, Croatian, Czech, Macedonian, Polish, Russian, Serbian, Slovak, Slovenian, Ukrainian;
- Other European (10): Albanian, Armenian, Esperanto, Estonian, Finnish, Georgian, Greek, Hungarian, Latvian, Lithuanian;
- Celtic/Irish (4): Breton, Irish, Scottish, Welsh;
- Turkic (5): Azerbaijani, Kazakh, Turkish, Tatar, Uzbek;
- Middle East (3): Arabic, Farsi, Hebrew;
- Niger-Congo/Afro-Asiatic (10): Amharic, Hausa, Igbo, Oromo, Somali, Swahili, Wolof, Xhosa, Yoruba, Zulu;
- Indo-Aryan (9): Bengali, Hindi, Marathi, Nepali, Oriya, Sinhala, Sindhi, Urdu, Tamil;
- Malayo-Polynesian (9): Cebuano, Ilocano, Indonesian, Javanese, Malagasy, Malay, Malayalam, Sundanese, Tagalog;
- Other Asian (2): Burmese, Khmer;

Most of these groups correspond to well established linguistic language families, but we have also performed some geographic groupings, in particular for small language families or isolated languages. We systematically mine all pairs within one language family. For instance, we provide bitexts for all pairs of Indian languages. In addition, we have identified major languages in each group and use them as *"bridge languages"* (underlined in the above list). We mine for all bitexts among these 27 bridge languages. The motivation for this bridge language approach is to connect the languages of the various groups, but sill avoid mining the full matrix.

We tried to support a large number of languages, but we are aware that the underlying LASER embedding is not very strong for all 90 languages, for instance several languages of the Niger-Congo family. Therefore, some of the mined bitexts may contain wrong alignments or even texts from other languages, despite careful filtering and two different LID classifiers. Additional cleaning/filtering may be needed. Nevertheless, we hope that these bitexts are a useful resources to support research in low-resource languages.

1.2 Example alignments

To illustrate the quality and richness of the mined bitexts, we provide here some examples of extracted bitexts. We first searched for English sentences which appear simultaneously in bitexts for ten different languages (Arabic, German, French, Indonesian, Japanese, Korean, Russian, Turkish Vietnamese and Chinese). About ten thousand sentences are such 11-way parallel. Table 1 gives four examples. The first two examples are very generic sentences which could appear on many Web pages. This nicely showcases the potential of global mining to find mutual translations in unrelated documents. The second example is rather long sentence from some political document. Finally, we provide an example from the medical domain. We also observe grammatically wrong sentences, e.g. *"Ein Besuch in einem kranken Freund"*. This may indicate that our approach can find parallel sentences which were translated by (low quality) MT.

En	You should clean the refrigerator once a month.	Visiting a sick friend.
Ar	وأخيرا تذكري أنه يجب عليكي تنظيف الثلاجة مرة واحدة في الشهر.	زرت صديقا مريضا.
De	Den Kühlschrank sollten Sie einmal im Monat saubermachen.	Ein Besuch in einem kranken Freund
Fr	Il est recommandé de nettoyer le réfrigérateur une fois par mois.	visite à un ami malade.
Id	Sebulan sekali kulkas harus dibersihkan.	Kunjungi teman yang sakit
Ja	１ヶ月に１回くらいは冷蔵庫の蔵ざらえをしなきゃ。	病の友達を訪ねる
Ko	한 달에 한 번 정도는 냉장고 청소를 해주는 게 좋다.	아픈 친구를 보는 심정으로
Ru	Холодильник следует размораживать раз в месяц.	Посещение больного друга.
Tr	Buzdolabını boşaltarak ayda bir kez temizleyin.	Hasta bir dostu ziyaret etmek.
Vi	Vì vậy, mỗi tháng bạn nên vệ sinh tủ lạnh một lần.	Thăm người bạn THÂN bệnh
Zh	如果有必要，你可以一个月清理一次冰箱。	探望一个生病的朋友。

En	With the growing importance of world trade and the global community, business executives and legal professionals are expected to look beyond national jurisdictions and understand issues of international law and international commercial law.
Ar	ومن المتوقع أن تنظر إلى أبعد السلطات القضائية الوطنية وفهم قضايا القانون الأوروبي والدولي المستشارين القانونيين، مع تزايد أهمية التجارة العالمية والمجتمع العالمي.
De	Da Handel und Unternehmen immer globaler werden, wird erwartet, dass Rechtsberater über nationale Zuständigkeiten hinausblicken und Fragen des europäischen und internationalen Rechts verstehen.
Fr	Avec l'importance croissante du commerce mondial et la communauté mondiale, consultants juridiques devraient regarder au-delà des juridictions nationales et de comprendre les questions de droit européen et international.
Id	Dengan semakin pentingnya perdagangan dunia dan masyarakat global, konsultan hukum diharapkan untuk melihat melampaui yurisdiksi nasional dan memahami masalah hukum Eropa dan internasional.
Ja	法律コンサルタントは、貿易とビジネスがますますグローバル化するにつれて、国の管轄権を超えて、欧州および国際法の問題を理解することが期待されています。
Ko	무역 및 비즈니스가 전 세계적으로 증가함에 따라 법률 컨설턴트는 국가 관할권을 넘어서서 유럽 및 국제법 문제를 이해할 것으로 예상됩니다.
Ru	С ростом важности мировой торговли и мирового сообщества, юридические консультанты, как ожидается, искать за пределами национальной юрисдикции и понимания вопросов европейского и международного права.
Tr	Ticaret ve iş dünyası gittikçe küreselleştikçe, hukuk müşavirlerinin ulusal yargıların ötesine geçmesi ve Avrupa ve uluslararası hukuk konularını anlamaları beklenmektedir.
Vi	Với tầm quan trọng ngày càng tăng của thương mại thế giới và cộng đồng quốc tế, tư vấn pháp luật được dự kiến để nhìn xa hơn khu vực pháp lý quốc gia và hiểu các vấn đề của pháp luật châu Âu và quốc tế.
Zh	随着世界贸易和全球社会的重要性日益增加，法律顾问有望超越国家管辖和了解欧洲和国际法律的问题。

En	When we breathe quickly we also build up oxygen in our blood.
Ar	عندما نتنفس بسرعة نقوم ببناء الأكسجين في دمائنا.
De	Wenn wir schnell atmen, bauen wir auch Sauerstoff in unserem Blut auf.
Fr	Lorsque nous respirons rapidement, nous créons également de l'oxygène dans notre sang.
Id	Ketika kita bernapas dengan cepat, kita juga membangun oksigen dalam darah kita.
Ja	私たちが素早く呼吸すると、血液中に酸素も蓄積します。
Ko	우리가 빨리 숨을 쉬면 우리도 피 속에 산소를 축적합니다.
Ru	Когда мы дышим быстро, мы также накапливаем кислород в нашей крови.
Tr	Khi chúng ta thở nhanh, chúng ta cũng tích tụ oxy trong máu.
Vi	Çabucak nefes aldığımızda, kanımızda da oksijen biriktiririz.
Zh	当我们快速呼吸时，我们的血液中也会积聚氧气。

Table 1: Examples of English sentences for which alignments in at least ten languages were found.

Model	Test set	de-en	en-de	en-ru	ru-en	zh-en	en-zh	de-fr	fr-de
Ott et al. (2018)	Newstest2014	-	28.6	-	-	-	-	-	-
Fan et al. (2019)	Newstest2014	-	29.6	-	-	-	-	-	-
CCMatrix	Newstest2012	31.6	25.3	-	-	-	-	-	-
CCMatrix	Newstest2013	34.9	29.3	30.2	32.5	-	-	-	-
CCMatrix	Newstest2014	38.9	32.2	45.7	43.8	-	-	-	-
CCMatrix	Newstest2015	38.2	34.4	38.4	37.8	-	-	-	-
CCMatrix	Newstest2016	46.6	40.7	36.8	37.9	-	-	-	-
CCMatrix	Newstest2017	40.2	32.9	41.0	43.1	30.4	37.5	-	-
CCMatrix	Newstest2018	49.9	50.3	35.7	36.9	30.2	40.8	-	-
CCMatrix	Newstest2019	43.3	44.5	35.5	41.8	34.8	35.6	37.9	33.5
CCMatrix	Newstest2020	39.2	35.1	25.5	37.1	35.0	38.8	33.8	33.8

Table 2: Detokenized SacreBLEU scores of CCMatrix models on all the available Newstest sets.

1.3 Additional WMT Results

We provide translation results as measured by SacreBLEU (Post, 2018) on all available WMT test sets in Table 2. For one of the most common translation evaluation benchmarks in the community, training on WMT16 en-de and evaluating on Newstest2014, we display the current state of the art results as well as the result of a well trained standard Transformer, to provide contrast against training on mined data only. On this benchmark, we find improvements of almost 2 BLEU points.

References

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling Neural Machine Translation. *arXiv:1806.00187*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.