MT Marathon in the Americas 2019 (College Park)

Proposed Projects

Feel free to start a new entry or add your comments anywhere, in the text or on side.

Projects can be proposed until the first day of MT Marathon, but announcing them earlier might attract more participants, come better prepared etc.

---

1. Finetuning for MT Robustness to Natural Noise
2. Recognizing misleading MT output
3. NMT for Query Translation in Cross-Language Search
4. Using typological information for true zero-resource NMT
5. Open/Competitive Evaluation Leaderboard of Translations (OCELoT)
6. An Exploration of Pre-trained Embedding for Better Domain-invariant Quality Estimation
7. Minimum Risk Training

Finetuning for MT Robustness to Natural Noise

Antonis Anastasopoulos

Shared doc:

https://docs.google.com/document/d/1HuCEodMu07JRdbSbwj5GIjxIK2AW2fmzWPCvZYyEI_g/edit?usp=sharing

As several works have noted, MT quality degrades when confronted with test-time source-side noise, either synthetic ("Synthetic and Natural Noise both Break NMT" Belinkov and Bisk, ICLR 2018) or natural, e.g. grammatical errors ("NMT of Text from Non-Native Speakers" Anastasopoulos et al, NAACL 2019). The main method for dealing with this is to synthesize noise on the source part of the parallel data, and train on the artificially noised data. Instead, we propose to use real data with naturally-occurring grammatical errors that non-native speakers make, using datasets from the Grammar Error Correction literature, which typically provide noisy and clean versions of the same sentence. The caveat is that this data does not come with translations, which are needed for training a MT system. A possible solution to this (hinted at "An Analysis of Source-Side Grammatical Errors in NMT", Anastasopoulos, BlackboxNLP 2019, to appear) is to treat the translation of the clean source sentence as the target for the noisy version. In this project we will start with high-quality pre-trained Eng->X NMT systems (e.g. from fairseq, openNMT, or others) and explore continued training for robustness with real source-side noisy/clean data, but without access to gold target sentences.

Other related work:

- "Towards Robust NMT", Cheng et al, ACL 2018
- WMT Robustness Challenge
- "Improving Robustness of Machine Translation with Synthetic Noise", Vhaibav et al, NAACL 2019
- "MTNT: A Testbed for Machine Translation of Noisy Text", Michel and Neubig, EMNLP 2018.

Recognizing Misleading NMT Output

Marianna Martindale

NMT quality is often impressive, but when it fails it can fail catastrophically. Sometimes this is in ways that are obvious to the user (e.g., repeating a word or phrase), but sometimes the output is fluent enough that the user might be misled into thinking the output is correct when the meaning is completely wrong. Recognizing these misleading translations is the first step in teaching the system not to make them and/or warning the user when they occur.

One approach is to score the segment translations for fluency and adequacy and label segments that are fluent but not adequate as potentially misleading. Previous work (to appear in MT Summit) explores some approaches, but there's plenty of room for improvement.

Related work:

- Why we should care about these errors even though they're rare: "**Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT**" (Martindale & Carpuat 2018); In 2016, Google Translate was translating 100 billion words per day. Assuming the current numbers are at least that high and an average sentence length of 10, if even 0.05% of sentence translations are misleading that's 50,000 misleading translations per day.
- WMT Metrics and Quality Estimation tasks. Note that although the metrics task objective is to match human *adequacy* judgments, disfluency seems to adversely affect these scores
- Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation (accepted MT Summit submission, not final version)

Available data:

- WMT16 results included some fluency direct assessment annotations (collected for accuracy at the system level, so few segments have more than two annotations). (cz-en, de-en, fi-en, ro-en, ru-en, tr-en)
- WMT metrics task results for 2016, 2017, and 2018 include segment-level adequacy direct assessment annotations (see caveat above about human adequacy judgments being influenced by fluency)
- We have a handful of segments annotated for fluency, plausibility, adequacy, and misleadingness:

Annotators per segment / Language	1+	2+	3+
Arabic-EN	409	197	88
Farsi-EN	1491	603	41
Korean-EN	965	570	397

Combined working notes document

NMT for Query Translation in Cross-Language Search

Petra Galuscakova, Suraj Nair, Doug Oard

In Cross-language Information Retrieval (CLIR) we have a collection of foreign language documents in which we are trying to search using queries in different language (e.g. English). CLIR systems typically translate the foreign documents into English and then apply standard information retrieval techniques on the translated documents. However translating the huge amounts of documents is slow and thus not always convenient or even possible. Another approach is thus to translate the queries into the language of the documents. However, the standard search queries differ from the sentences found in typical MT training corpora -- e.g. they often consist of just a few words, are not always fluent and do not use punctuation, and rare words carry salient relevance information. In this project, we would like to evaluate various techniques for tailoring NMT to query translation.

The project will consists of three main tasks:

1) **Building a collection** for training MT system for query translation. We would like to build an artificial collection of the queries from either a) Wikipedia or b) parallel data.
    a) Obtain Wikipedia titles in different languages using the inter-language links.
    b) In the case of using the parallel data, the co-occurrent word n-grams will be randomly selected from the English side and their alignments (acquired by GIZA++) from the foreign side will be found. The pair of n-grams and their alignments will be then used as the training instances.
2) **Building MT system** which will be trained on the created collection using and adapting existing NMT toolkits (e.g. MARIAN, Sockeye, OpenNMT).
3) **Evaluation**. As BLEU score is not expected to be helpful for evaluating the task due to the lack of fluency, different MT evaluation measure need to be used (e.g. unigram precision/recall or PER) or the system will be evaluated extrinsically using the effect of MT on the search quality (for example on the CLEF collection).

Relevant papers:

- Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth J.F. Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, Martin Popel, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová: Adaptation of machine translation for multilingual information retrieval in the medical domain , In: Artificial Intelligence in Medicine, Volume 61, Issue 3, July 2014, Pages 165-185
  https://www.sciencedirect.com/science/article/pii/S0933365714000062
- Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. ACM Comput. Surv. 45, 1, 2011
  https://dl.acm.org/citation.cfm?id=2379777

Data:

- Khresmoi Query Translation Test Data can be possibly used for the evaluation:
  http://hdl.handle.net/11234/1-2121
  https://ufal.mff.cuni.cz/~pecina/files/lrec-2014.pdf
- CLEF 2000-2003: http://catalog.elra.info/product_info.php?products_id=888

Using typological information for true zero-resource NMT

Antonis Anastasopoulos

Shared doc:

https://docs.google.com/document/d/1eSBQVFJCw8DREzEzqUNcJ35WqViy0j4Oj6pNn3mLcC0/edit?usp=sharing

Most zero-shot approaches rely on the existence of some data in the languages of interest, except not parallel in the desired directions. For example, one can train a multilingual system on Eng<-->{Fra,Deu}, and then attempt Fra->Deu translation in a zero-shot fashion. Often, language embeddings for the source and target languages are learned (e.g. Wu et al 2016). Other works rely on mapping the embedding spaces between languages. For many low-resource languages (or dialects), the only available data are monolingual, and hence not suitable, as they cannot be readily used for learning these language embeddings. What we propose is to use pre-computed typological features, which have been shown to be useful in e.g. Language Modeling.

We will use a portion of TED data and train a multilingual {Ces,Por,Rus,Tur}->Eng system, also utilizing the typological features from URIEL (see the lang2vec python package), and then attempt to do *true* zero-shot MT for the related {Slk,Glg,Bel,Aze}->Eng.

(These languages are chosen to match previous work for easy comparisons -- we could expand to more or other datasets).

Some (non multilingual, not using any typological features) baseline zero-shot results:

Train : X to English	Test: X to English	BLEU
Czech	Slovak	8.61
Portuguese	Galician	12.37
Spanish	Galician	10.71
Turkish	Azerbaijani	4.29
Serbian	Bosnian	29.21

Open/Competitive Evaluation Leaderboard of Translations (OCELoT)

Christian Federmann <chrife@microsoft.com>

TL;DR: Let's build an open platform for competitive machine translation quality tracking!
https://GitHub.com/cfedermann/OCELoT

Research progress in several machine learning disciplines is tracked via open/competitive leaderboards. While MT has this on a yearly cadence, via the WMT Conference for Machine Translation shared tasks, there is no continuous competition and teams not participating in WMT are left out…

The goal of this project is to build an openly accessible NMT leaderboard where anybody can submit their translation output, both for automatic scoring via SacreBLEU, but also (on some sensible cadence, for the top-n systems to control annotation costs) for human evaluation, adopting WMT's methodology.

You can join this project by 1) contributing code (focusing on Python 3, Django 2, TDD, pylint and Black, amongst others), or by 2) breaking our prototype via submissions of your NMT systems' output. Either will be helpful and much appreciated!

I'll find some human annotation budget for our first batch of participating systems! If this project triggers enough interest from the community, we can look into keeping this alive...

Looking forward to your feedback and questions at MTMA19!

An Exploration of Pre-trained Embedding for Better Domain-invariant Quality Estimation

Shuoyang Ding <dings@jhu.edu> Nanyun Peng <nanyunpe@usc.edu>

Most existing sentence-level quality estimation model relies on either a linear classifier or a full encoder-decoder neural MT architecture trained for accurate prediction of HTER (e.g. the

winning UNQE and QEBrain system in WMT quality estimation shared task 2018). This worked pretty well, but is expensive, and carries the same domain-adaptation problem as any classifier/NMT model would do.

Alternatively, QE could be positioned as a MT evaluation metric, but computed between source input and system output (instead of system output and reference). However, since exact string match between source and system output is not possible, such metric would only be feasible if some kind of semantic representation (e.g. word embedding) is available.

Unfortunately, for a long time, directly using pre-trained word embedding was not proved to be very helpful for segment-level MT evaluation (https://www.aclweb.org/anthology/W16-4505). Supervised metric, such as BLEND (https://www.statmt.org/wmt17/pdf/WMT68.pdf) and RUSE (https://www.aclweb.org/anthology/W18-6456) works better, but require human evaluation data to train on, which raises the same domain adaptation problem as UNQE etc..

However, a very recent paper on MT evaluation metric (https://arxiv.org/pdf/1904.09675.pdf) proposed a very interesting method based on BERT[1] that does not require any task-specific tuning to beat BLEND and RUSE, which might indicate that the advancements in pre-trained word embedding is large enough to push the boundary of metric research.

We propose to examine how far applying similar method with multilingual BERT can get us in quality estimation task. Basically, here is what we intend to do (in the order of exploration):

1. Hack the code of BERTScore to run QE on some WMT QE task data.
2. How domain-specific is this? Need some idea on how to evaluate this. In the worst case, we'll have to do a little bit human annotation on-the-fly.
3. How does BERTScore compare with simple cosine similarity given by pre-trained multilingual *sentence* embeddings, such as LASER?
4. BERTScore induce alignment between translation and reference by maximizing word similarity. This might be good enough for MT evaluation, but that might potentially be a bigger problem for QE? How much does alignment quality matter for QE in this scenario?
5. Beyond QE, is source sentence useful for MT evaluation as well? Presumably, the source sentence could also give the humans some idea about adequacy, but this might be hard to evaluate. Maybe the adequacy/fluency human annotation from WMT is the way to go?

---

[1] Basically, it's a tf-idf-weighted unigram similarity, where the unigram similarity is given by the cosine distance between the reference and candidate words. The alignment between the reference and candidate is constructed by greedily maximizing the similarity scores for each candidate/reference words, depending on whether precision or recall is computed.

Combined working notes document