MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora

**Motivations**
- Multiword Expressions (MWEs) set challenges for state-of-the-art neural machine translation (NMT).
- Bilingual parallel corpora introducing MWE annotations are very scarce.
- How to extract automatically: Manual annotation is costly and time-consuming.

**Strategies**
- Automatic extraction of bilingual MWEs from parallel corpora introducing new language pairs en-de/zh.
- Data augmentation for MT system training using extracted bilingual MWEs as Knowledge Base.
- Quantitative and qualitative evaluation of MT system outputs looking into MWEs.

**Experimental Settings and Outcomes**
- Transformer models learned from scratch using 5 million parallel sentences from WMT.
- Our collections are 3,159,226 and 143,042 bilingual MWE pairs for German-English and Chinese-English respectively after filtering.
- We examine the quality of these extracted bilingual MWEs in MT experiments.
- Our initial experiments applying MWEs in MT show improved translation performances on MWE terms in qualitative analysis and better general evaluation scores in quantitative analysis, on both German-English and Chinese-English language pairs. (Figure examples => right side)

**Procedures for MWE extraction**
- Morphological tagging of De/Zh and En.
- Tagged De/Zh/En into XML format.
- Design MWE-patterns for De/Zh/En
- Extract Monolingual MWEs with MWEtoolkit
- Generate De/Zh-En lexicon translation probability files with Giza++ and Moses
- Align Bilingual MWEs with MPAligner

**Sample Comparisons of MT Systems**

| Examples of MWE translations in MT outputs |
|------------------------------------------|
| **Src** | 俄罗斯与土耳其领导人周二进行会见，双方握手并宣布正式结束长达八个月的口水战与经济制裁。 |
| **Ref** | the leaders of Russia and Turkey met on Tuesday to shake hands and declare a formal end to an eight-month long war of words and economic sanctions. |
| **Base** | Russian and Turkish leaders met Tuesday, shaking hands and declaring the official end of eight months of water fighting and economic sanctions. |
| **B+MWE** | Russian and Turkish leaders met on Tuesday and both shook hands and announced a formal end of eight months of oral combat and economic sanctions. |

**MultiMWE corpora creation workflow**

![MultiMWE workflow diagram]

- **Transformer models learned from scratch using 5 million parallel sentences from WMT**
- **Our collections are 3,159,226 and 143,042 bilingual MWE pairs for German-English and Chinese-English respectively after filtering.**
- **We examine the quality of these extracted bilingual MWEs in MT experiments.**
- **Our initial experiments applying MWEs in MT show improved translation performances on MWE terms in qualitative analysis and better general evaluation scores in quantitative analysis, on both German-English and Chinese-English language pairs. (Figure examples => right side)**

**Procedures for MWE extraction**
- Morphological tagging of De/Zh and En.
- Tagged De/Zh/En into XML format.
- Design MWE-patterns for De/Zh/En
- Extract Monolingual MWEs with MWEtoolkit
- Generate De/Zh-En lexicon translation probability files with Giza++ and Moses
- Align Bilingual MWEs with MPAligner

**Sample Comparisons of MT Systems**

| Examples of MWE translations in MT outputs |
|------------------------------------------|
| **Src** | 俄罗斯与土耳其领导人周二进行会见，双方握手并宣布正式结束长达八个月的口水战与经济制裁。 |
| **Ref** | the leaders of Russia and Turkey met on Tuesday to shake hands and declare a formal end to an eight-month long war of words and economic sanctions. |
| **Base** | Russian and Turkish leaders met Tuesday, shaking hands and declaring the official end of eight months of water fighting and economic sanctions. |
| **B+MWE** | Russian and Turkish leaders met on Tuesday and both shook hands and announced a formal end of eight months of oral combat and economic sanctions. |

**Figure:** Baseline model vs model with filtered MWEs integrated