XL-AMR: Enabling Cross-Lingual AMR Parsing with Transfer Learning Techniques

Rexhina Blloshmi **Rocco Tripodi** **Roberto Navigli**
Sapienza NLP Group
Department of Computer Science, Sapienza University of Rome
{blloshmi,tripodi,navigli}@di.uniroma1.it

Abstract

Abstract Meaning Representation (AMR) is a popular formalism of natural language that represents the meaning of a sentence as a semantic graph. It is agnostic about how to derive meanings from strings and for this reason it lends itself well to the encoding of semantics across languages. However, cross-lingual AMR parsing is a hard task, because training data are scarce in languages other than English and the existing English AMR parsers are not directly suited to being used in a cross-lingual setting. In this work we tackle these two problems so as to enable cross-lingual AMR parsing: we explore different transfer learning techniques for producing automatic AMR annotations across languages and develop a cross-lingual AMR parser, XL-AMR. This can be trained on the produced data and does not rely on AMR aligners or *source-copy* mechanisms as is commonly the case in English AMR parsing. The results of XL-AMR significantly surpass those previously reported in Chinese, German, Italian and Spanish. Finally we provide a qualitative analysis which sheds light on the suitability of AMR across languages. We release XL-AMR at github.com/SapienzaNLP/xl-amr.

1 Introduction

Abstract Meaning Representation (AMR) is a popular formalism for natural language (Banarescu et al., 2013). It represents sentences as rooted, directed and acyclic graphs in which nodes are concepts and edges are semantic relations among them. AMR unifies, in a single structure, a rich set of information coming from different tasks, such as Named Entity Recognition (NER), Semantic Role Labeling (SRL), Word Sense Disambiguation (WSD) and coreference resolution. Such representations are actively integrated in several Natural Language Processing (NLP) applications, *inter alia*, information extraction (Rao et al., 2017), text summarization (Hardy and Vlachos, 2018; Liao et al., 2018), paraphrase detection (Issa et al., 2018), spoken language understanding (Damonte et al., 2019), machine translation (Song et al., 2019b) and human-robot interaction (Bonial et al., 2020). It is therefore desirable to extend AMR semantic representations across languages along the lines of cross-lingual representations for grammatical annotation (de Marneffe et al., 2014), concepts (Conia and Navigli, 2020) and semantic roles (Akbik et al., 2015; Di Fabio et al., 2019). Furthermore, it could be especially useful to integrate cross-lingual semantic structures in multilingual applications of natural language understanding.

A peculiar feature of the AMR formalism is that it aims at abstracting away from word forms. AMR graphs are *unanchored*, i.e., the linkage between tokens in a sentence and nodes in the corresponding graph is not explicitly annotated. Hence, the feature of being agnostic about how to derive meanings from strings makes AMR particularly suitable for representing semantics cross-lingually. However, AMR was initially designed for encoding the meaning of English sentences. Owing to this, the available resources and modelling techniques focus mostly on English, while leaving cross-lingual AMR understudied. Some preliminary studies showed the limits of AMR as an interlingua, categorizing them as due to distinctions in the underlying ontologies or structural divergences among languages (Xue et al., 2014; Hajič et al., 2014). More recent studies, instead, have provided evidence that AMR or a simplified version of it can be used as a formalism for cross-lingual semantic representation, showing that it is possible to overcome some of the structural linguistic divergences (Damonte and Cohen, 2018; Zhu et al., 2019).

The underlying idea of this paper is that AMR can be used to represent semantic information in

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2487–2500,
November 16–20, 2020. ©2020 Association for Computational Linguistics

different languages since there exist key linguistic features that are shared across languages, such as predicates, roles and conjunctions (Von Fintel and Matthewson, 2008). However, developing an AMR parser for multiple languages is hard because the existing annotated training resources that are sufficiently large are available in English only, and, moreover, acquiring semantic annotations for a large number of sentences is well-known to be a slow and expensive process in NLP (Zhang et al., 2018; Pasini, 2020). To this end, we aim at exploiting and developing the necessary tools and resources for enabling cross-lingual AMR parsing, i.e., the task of transducing a sentence in the source language into an AMR graph based on English (Damonte and Cohen, 2018).

We present XL-AMR, a cross-lingual AMR parser, and study different transfer learning techniques to enable its training: i) model transfer which relies on language-independent features, ii) annotation projection relying on parallel corpora and available English AMR parsers, and iii) automatic translation of the training corpora which guarantees gold AMR structures. We make the following contributions:

- We develop and release XL-AMR, a cross-lingual AMR parser which disposes of word aligners, i.e., word-to-word and word-to-node, and surpasses the previously reported results on Chinese, German, Italian and Spanish, by a large margin.
- Exploration of different techniques to create cross-lingual AMR training data, showing how it is possible to transfer semantic structure information across different languages.
- Creation and release of diverse quality silver data for cross-lingual AMR parsing.
- Qualitative analysis of the ability of XL-AMR to transfer semantic structures across languages and of AMR to represent the meaning of sentences cross-lingually.

2 Related Work

Our work lies between two areas, namely, semantic parsing and cross-lingual transfer learning.

Semantic parsing Semantic parsing is a key task required to complete the puzzle of Natural Language Understanding (Navigli, 2018), and one which is receiving growing attention in the scientific community. Besides AMR, various different formalisms have been proposed over the years to encode semantic structures: Elementary Dependency Structures (Oepen and Lønning, 2006, EDS), Prague Tectogrammatical Graphs (Hajič et al., 2012, PTG), Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013, UCCA), Universal Decompositional Semantics (White et al., 2016, UDS), *inter alia*. While some frameworks, such as UCCA and UDS, have been exploited in a cross-linguistic setting (Lyu et al., 2019; Zhang et al., 2018), cross-lingual AMR has mainly been studied within the scope of annotation analysis works (Xue et al., 2014; Hajič et al., 2014). These works point out the limitations of AMR as an interlingua, and consider them partly due to the distinctions in the underlying ontologies and structural divergences among languages. Zhu et al. (2019) also evaluate the properties of AMR across languages and aim at simplifying this formalism in order to express only essential semantic features of a sentence, such as predicate roles and linguistic relations. Cross-lingual AMR *parsing*, instead, has received relatively less attention. This is largely attributable to the lack of training data and evaluation benchmarks in languages other than English. Damonte and Cohen (2018) propose, to the best of our knowledge, the only cross-lingual AMR parser to date and, moreover, their proposed cross-lingual AMR evaluation benchmark has been released only very recently (Damonte and Cohen, 2020). The authors adapt a transition-based English AMR parser (Damonte et al., 2017) for cross-lingual AMR parsing, which is trained on silver annotated data. However, the performances it has achieved are not satisfying in terms of Smatch score (Cai and Knight, 2013), mostly as a result of concept identification errors, which in turn are directly related to the usage of noisy word-to-node alignments projected from English. Throughout the literature English AMR parsers commonly rely on AMR alignments which are automatically created using heuristics (Flanigan et al., 2014), or on pretrained aligners (Pourdamghani et al., 2014; Liu et al., 2018), treated as latent variables of the model (Lyu and Titov, 2018) or implicitly modelled through *source-copy* mechanisms (Zhang et al., 2019). These alignments, however, take advantage of the fact that AMR nodes and English words are highly related.[1] This dependency is therefore not suitable for cross-lingual parsing since similarity between words in

[1] In AMR 2.0 roughly 60% of the nodes are English words. In addition, PropBank predicates are often similar to English words, e.g., one can heuristically align *publish-01* to *publish*.

the sentences and concepts in the graph does not hold at large. Our parser, instead, disposes of explicit and implicit AMR alignments using a *seq2seq* model for concept identification and achieves significantly higher performance on all the tested languages. On the other hand, to account for data sparsity, XL-AMR employs several common techniques in English AMR parsing literature (Konstas et al., 2017; Zhang et al., 2019), such as anonymization and recategorization, expanding them across languages by relying on multilingual resources.

Transfer learning The idea behind this method is to leverage annotations available in one language, commonly English, to enable learning models that generalize to languages where labelled resources are scarce (Ruder et al., 2019). Different techniques include annotation projection, machine translation and language-independent feature-based models. Extensive works in this direction exist, applied to different NLP tasks, i.e., WSD (Barba et al., 2020), SRL (Padó and Lapata, 2009; Kozhevnikov and Titov, 2013), Dependency Parsing (Tiedemann, 2015), concept representation (Conia and Navigli, 2020), etc. In cross-lingual AMR parsing, annotation projection is employed by Damonte and Cohen (2018), who produce cross-lingual silver AMR annotations by exploiting parallel sentences selected from the Europarl corpus (Koehn, 2005): English sentences are parsed using an English parser (Damonte et al., 2017, AMREAGER) and the resulting graphs are associated with the corresponding parallel sentences. However, the data on which AMREAGER was trained is very different from those used to produce the silver annotations, thus affecting the quality and reliability of the AMR graphs produced. Here we test two different techniques: we conduct experiments with annotation projection using Europarl for comparison, and, in addition, we use translation techniques to produce better quality training corpora. This leads to significant improvements and provides evidence that better quality data – and models – allow for using AMR as an interlingua.

3 Cross-Lingual AMR

In what follows we first formalize the task (Section 3.1) and then detail our cross-lingual AMR parser (Section 3.2) and our proposed silver data creation methods (Section 3.3). Finally, we list the pre- and postprocessing cross-lingual techniques and resources we employ (Section 3.4).

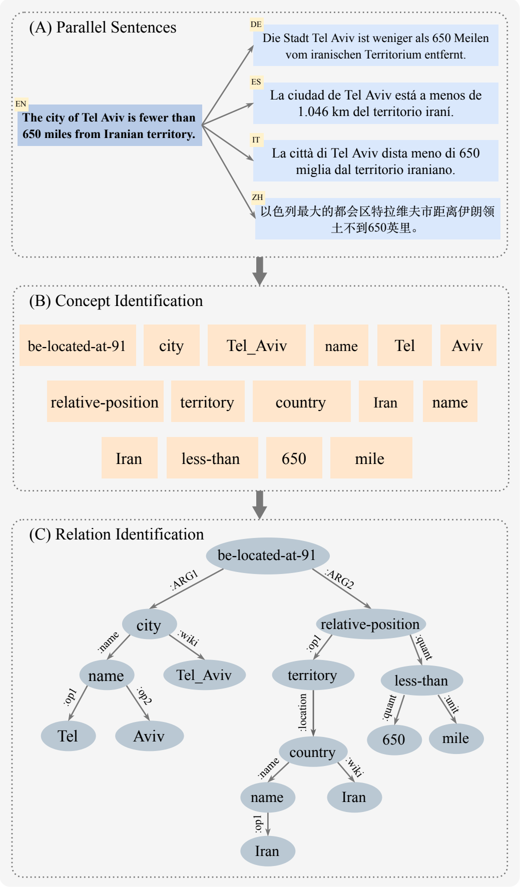
(A) Parallel Sentences
EN
The city of Tel Aviv is fewer than 650 miles from Iranian territory.
DE
Die Stadt Tel Aviv ist weniger als 650 Meilen vom iranischen Territorium entfernt.
ES
La ciudad de Tel Aviv está a menos de 1.046 km del territorio iraní.
IT
La città di Tel Aviv dista meno di 650 miglia dal territorio iraniano.
ZH
以色列最大的都会区特拉维夫市距离伊朗领土不到650英里。
(B) Concept Identification
be-located-at-91
city
Tel_Aviv
name
Tel
Aviv
relative-position
territory
country
Iran
name
Iran
less-than
650
mile
(C) Relation Identification
be-located-at-91
:ARG1
:ARG2
city
:name
:wiki
name
Tel_Aviv
:op1
:op2
Tel
Aviv
relative-position
:op1
:quant
territory
less-than
:location
:quant
:unit
650
mile
country
:name
:wiki
name
Iran
:op1
Iran

Figure 1: Cross-Lingual AMR Parsing: (A) Sentences written in different languages sharing the same meaning; (B) concepts representing the words in the sentences; (C) the final AMR graph.

3.1 The Task

Cross-lingual AMR parsing is defined as the task of transducing a sentence in any language to the AMR graph of its English translation whose nodes are either English words, PropBank framesets (Kingsbury and Palmer, 2002) or special AMR keywords.

Breaking down this definition, given an English sentence and its translation T_L in a language L, their meaning representation is ideally formalized by the same AMR, $G = (V, E)$, where V is a list of concept nodes and E is the set of semantic relations between them. Figure 1-A shows an example of a sentence in English, with its translations into Chinese, German, Italian and Spanish which have the same meaning and therefore the same abstract representation (Figure 1-C). Following state-of-the-art models for English AMR parsing (Zhang et al., 2019), we tackle cross-lingual AMR parsing as a two-stage approach, i.e., concept and relation identification, which we briefly overview here and later detail in Section 3.2. For concept identification, given the sequence $T_L = (t_1, t_2, \ldots, t_j)$, t_i being a word in language L ($i \in \{1, \ldots, j\}$, $L \in \{\text{EN, DE, ES, IT, ZH}\}$), we train a neural network to generate the list of nodes $V = (v_1, v_2, \ldots, v_n)$, $v_i \in$ English words $\cup$ PropBank framesets $\cup$ AMR keywords. In Figure 1-B we show the list of concepts that represent the words in the sentences of Figure 1-A. The relation identification procedure, instead, is inspired by the arc-factored approaches employed in dependency parsing (Kiperwasser and Goldberg, 2016), i.e., searching for the maximum-scoring connected subgraph over the identified concepts in the previous step. Thus, given the list of predicted nodes $V = (v_1, v_2, \ldots, v_n)$ and a learned score for each candidate edge, we search for the highest-scoring spanning tree and then merge the duplicate nodes based on unique node indices (see Section 3.2) to restore the final AMR graph. Figure 1-C shows the AMR representing the shared semantics of the sentences in Figure 1-A.

3.2 XL-AMR Model

XL-AMR is composed of two modules which are learned jointly, i.e., concept identification, modeled as a *seq2seq* problem, and relation identification, based on a biaffine attention classifier (Dozat and Manning, 2017). We use a *seq2seq* model to dispose of the need for an AMR alignment module. Lyu and Titov (2018) argue that alignments are important for injecting a useful inductive bias for AMR parsing and maintain that alignment-based parsers might be better than *seq2seq* for AMR parsing, owing to the relatively small amount of data available for AMR. However, aligning words to AMR nodes in cross-lingual parsing is challenging. The widely used AMR aligners are usually based on heuristics (Flanigan et al., 2014), or on the fact that AMR and English are highly cognate (Pourdamghani et al., 2014). Hence, these approaches would not be valid for cross-lingual alignment and, moreover, projecting the alignments across languages through English has shown to be noisy and to affect the parsing performance (Damonte and Cohen, 2018).

Concept identification At training time we obtain the list of nodes by first converting the graph into a tree, duplicating the nodes occurring in multiple relations, and then using a pre-order traversal over the tree. To account for reentrancies we assign a unique index to each node during traversal, similarly to Zhang et al. (2019). Following the attention-based encoder-decoder architecture proposed by Bahdanau et al. (2015), our concept identification module consists of a bidirectional RNN encoder and a decoder that attends to the source sentence at each concept decoding step.

The *encoder* employs an L-layer bidirectional RNN (Schuster and Paliwal, 1997) with LSTM cells (Hochreiter and Schmidhuber, 1997), i.e., BiLSTM, which encodes the input token embeddings e_i into hidden states h_i. Each hidden state $h_i^l = [\overrightarrow{h_i^l}; \overleftarrow{h_i^l}]$, is a concatenation of the forward hidden state and the backward hidden state at timestep i. Similarly to Zhang et al. (2019), the input token embedding e_i is a concatenation of contextualized embeddings, word embeddings, Part-of-Speech (PoS) embeddings, token anonymization indicator[2] and character-level embeddings. The subsequent BiLSTM layer, instead, takes the hidden states of the previous layer as input.

The *decoder* also consists of L recurrent neural network (unidirectional) layers with LSTM cells. The decoder embedding layer concatenates word embeddings, node index embeddings and character-level embeddings. The layer l of the decoder calculates $d_t^l = decoder_l(d_t^{l-1}, d_{t-1}^l)$, where d_t^{l-1} is the concept hidden state of the previous layer at timestep t while d_{t-1}^l that of previous timestep. d_0^l is initialized with the concatenation of the encoder's last hidden states $h^l = [\overrightarrow{h^l}; \overleftarrow{h^l}]$. We follow the *input feeding* approach of Luong et al. (2015), which concatenates the output of the decoder's embedding layer and an attentional vector computed at the previous timestep. We first compute the *source attention distribution* a_t using additive attention (Bahdanau et al., 2015) as follows:

$$e_{t,i} = v^\top \tanh(W_h h_i^L + W_s d_t^L + b_s)$$
$$a_t = \mathrm{softmax}(e_t)$$
$$c_t = \sum_i a_{t,i} h_i$$

where v, W_h, W_s and b_s are model parameters, and c_t is the source context vector. Then, we compute the attentional vector, $\tilde{d}_t = \tanh(W_c[c_t; d_t^L] + b_c)$, where W_c and b_c are model parameters.

Zhang et al. (2019) used the attentional vector to allow the decoder to copy nodes predicted in the previous steps (*target-copy*), rather than only

[2] Tokens representing named entities are anonymized during preprocessing and restored in postprocessing (Section 3.4).

generating a new node from the vocabulary. As they provide empirical evidence that this is crucial for handling reentrancies, we employ their *target-copy* approach and use the attentional vector $\tilde{d}_t$ to i) feed in a dense layer and softmax to produce a probability distribution over the vocabulary $P_{vocab} = \text{softmax}(W_{vocab}\tilde{d}_t + b_{vocab})$, ii) to learn a *target attention distribution* $\hat{a}_t$ (similar to the *source attention distribution* above), iii) to calculate p_{copy} and $p_{generate}$ probabilities that decide either to copy one of the previously predicted nodes by sampling a node from the *target attention distribution* $\hat{a}_t$, or to generate a new node from the output vocabulary. Each newly generated node is assigned a unique index, or it is assigned the index of the node copied from the previously generated concepts. At prediction time, we employ a beam search to decode the list of nodes based on the probability distribution computed above.

Relation identification For this module, we follow Zhang et al. (2019) and use a deep biaffine classifier inspired by Dozat and Manning (2017), which takes as input the decoder states and factorizes the edge prediction in two components predicting i) whether there is an edge between a pair of nodes, and ii) the edge label for each possible edge, respectively. We direct the reader to Zhang et al. (2019) and Dozat and Manning (2017) for technical details on the biaffine attention classifier. At prediction time, to ensure the validity of the tree, given the list of predicted nodes and the score for candidate edges, we search for the highest-scoring spanning tree using the Chu-Liu-Edmonds algorithm. We then merge the duplicate nodes based on the node indices to restore the final AMR graph.
The model is trained to jointly minimize the loss of reference nodes and edges.

3.3 Silver Training Data

In order to train cross-lingual AMR parsers and to evaluate the cross-lingual properties of AMR as an interlingua, we project existing AMR annotations for English sentences to target language sentences following two different approaches.

Parallel sentences - silver AMR graphs We follow Damonte and Cohen (2018) and project AMR graphs from English sentences to target language sentences through a parallel corpus. Differently from Damonte and Cohen (2018), we do not need *word-to-word* and *word-to-node* aligners for training the concept identification module. Instead we directly pair a sentence in the target language with the AMR graph corresponding to its English counterpart. In this case, while the sentences are parallel, the AMR graphs are of silver standard quality, i.e., the English sentences of the parallel corpus are parsed using an existing AMR parser. We refer to this method as PARSENTS-SILVERAMR.

Gold AMR graphs - silver translations In addition to pivoting through parallel sentences, we investigate whether considering human-annotated AMR graphs could bring more benefits than system produced AMR graphs. To this end, we make use of the existing gold standard datasets for AMR parsing, i.e., English sentence-AMR graph pairs, and use machine translation systems to translate the training sentences into the target language. This choice is motivated by the existence of reliable machine translation systems for the languages of our interest. Moreover, we validate the silver translations through a back-translation step (Sennrich et al., 2016). That is, firstly, we translate the sentences from English to the target language and, secondly, using the same neural translation model, we translate the target language translations back to English. Then, to filter out less accurate translations we apply a 1-NN strategy based on the cosine similarity between translations and source sentence semantic embeddings, similarly to Artetxe and Schwenk (2019a). If the nearest neighbour of a translation corresponds to its source English sentence, we consider it a good translation, otherwise we discard it. We employ semantic similarity since we have a two-step automatic translation, due to which lexical differences are introduced into translations compared to the original sentence. Typical machine translation metrics, e.g., BLEU, METEOR, rely on lexical similarity, which could lead good translations being discarded. In fact, we do not need the translation to be word-to-word aligned, but rather to preserve the meaning of the sentence, thus considering valid also the cases when certain words are translated into synonyms or related words. We refer to this method as GOLDAMR-SILVERTRNS.

3.4 Pre- and Postprocessing

AMR parsers in the literature rely on several pre- and postprocessing rules. We extend these rules for the cross-lingual AMR parsing task based on several multilingual resources such as Wikipedia, BabelNet 4.0 (Navigli and Ponzetto, 2010), DBpedia Spotlight API (Daiber et al., 2013) for wikifi-

Dataset	Lang	Train Insts	Dev Insts	Source
Gold	EN	36521	1368	AMR 2.0
ParSents SilverAmr	DE	20000	2000	Europarl
	EN	20000	2000	Europarl
	ES	20000	2000	Europarl
	IT	20000	2000	Europarl
GoldAmr SilverTrns	DE	34415	1319	AMR 2.0
	ES	34552	1325	AMR 2.0
	IT	34521	1322	AMR 2.0
	ZH	32154	1276	AMR 2.0

Table 1: Dataset quality standard, instances per language, and the source corpus of the sentences.

cation in all languages but Chinese, for which we use Babelfy (Moro et al., 2014) instead, Stanford CoreNLP (Manning et al., 2014) for English preprocessing pipeline, the Stanza Toolkit (Qi et al., 2020) for Chinese, German and Spanish sentences, and Tint[3] (Aprosio and Moretti, 2016) for Italian.

The preprocessing steps consist of: i) lemmatization, ii) PoS tagging, iii) NER, iv) re-categorization of entities and senses, v) removal of wiki links and polarity attributes. The postprocessing steps consist of restoring i) anonymized subgraphs, ii) wikification, iii) senses, iv) polarity attributes. We give full details on pre- and postprocessing in Appendix A.

4 Experiments

We now present a set of experiments for cross-lingual AMR parsing when using different training techniques and the silver data we created (see Section 3.3). We discuss the results of our multiple settings and compare with previous approaches performing cross-lingual AMR parsing.

Test bed We evaluate on the *Abstract Meaning Representation 2.0 - Four Translations* (Damonte and Cohen, 2020), a corpus containing translations of the test split of 1371 sentences from the LDC2017T10 (AMR 2.0), in Chinese (ZH), German (DE), Italian (IT) and Spanish (ES). This data is designed for use in cross-lingual AMR parsing (available to all LDC subscribers).

Dataset In Section 3.3, we explained the two projection approaches for obtaining cross-lingual AMR data, i.e., ParSents-SilverAmr and GoldAmr-SilverTrns.

For the first approach, inspired by Damonte and Cohen (2018), and for comparison purposes, we choose Europarl as parallel corpus.[4] We predict the silver AMR using the model of Zhang et al. (2019).

For the second approach, instead, i.e., GoldAmr-SilverTrns, we choose AMR 2.0 as gold dataset and translate the sentences into Chinese, German, Italian and Spanish. For German, Italian and Spanish, for both translating and back-translating the sentences we use the machine translation models made available by Tiedemann and Thottingal (2020, OPUS-MT).[5] For Chinese, instead, since OPUS-MT does not provide translation models, we employ the released MASS[6] (Song et al., 2019a) supervised neural translation models. Then, to filter out less accurate translations, we compute the cosine similarity between dense semantic representations of the original English sentence and its back-translated counterpart. To embed the sentences we use LASER (Artetxe and Schwenk, 2019b), a state-of-the-art model for sentence embeddings. Details on the number of instances per language and for each silver data approach are shown in Table 1.

Training configurations We conduct experiments following different training approaches:

- **Zero-shot** – the model is trained on English sentences only, relying on multilingual features, and is evaluated on all the target languages (henceforth ∅-shot).
- **Language-specific** – the model is trained only on target language data, i.e., DE, ES, IT or ZH, and evaluated in the same language.
- **Bilingual** – the model is trained on English data and one of either DE, ES, IT or ZH, and evaluated in the target language.
- **Multilingual** – the model is trained on data from all available languages per setting and evaluated on the target languages.

Systems We denote the variations of XL-AMR, based on the above training configurations, as XL-AMR^{data} where, $data \in \{par, trans, amr\}$, par referring to the data produced with ParSents-SilverAmr approach, $trans$ to GoldAmr-SilverTrns approach, amr to the AMR 2.0 English gold standard, and $data$+ refers to combining par or $trans$ with amr. The only existing cross-lingual AMR parser from the literature to date is

[3] Stanza does not provide a NER model for Italian.

[4] We do not produce silver AMR graphs for Chinese since Europarl does not cover the Chinese language.

[5] We provide the list of models we used in Appendix B.

[6] github.com/microsoft/MASS

Parser	Configuration	DE	ES	IT	ZH
AMREAGER	Lang-Spec.	39.0	42.0	43.0	35.0
XL-AMR$_{\emptyset}^{amr}$	$\emptyset$-shot	32.7	39.1	37.1	25.9
XL-AMR$_{\emptyset}^{par+}$	$\emptyset$-shot	38.3	41.8	41.0	23.9
XL-AMR^{par}	Lang-Spec.	40.8	44.2	43.4	-
	Multiling.	41.5	45.6	45.0	-
	Biling.	42.7	47.9	46.7	-
XL-AMR^{par+}	Multiling.	46.3	51.2	50.9	-
	Biling.	47.0	53.0	51.4	-
XL-AMR^{trans}	Lang-Spec.	51.6	56.1	56.7	**43.1**
	Multiling.	49.9	53.0	54.0	40.0
	Multiling. (-ZH)	51.5	55.5	55.9	-
XL-AMR^{trans+}	Multiling.	49.9	53.2	53.5	41.0
	Multiling. (-ZH)	52.1	56.2	56.7	-
	Biling.	**53.0**	**58.0**	**58.1**	41.5

Table 2: Smatch F1 scores on DE, ES, IT and ZH. Best scores per language are denoted in **bold**.

the one of Damonte and Cohen (2018, AMREAGER Multilingual), henceforth AMREAGER. We compare the results of the XL-AMR variants with the projection method of AMREAGER on the gold dataset, i.e., AMR *2.0 - Four Translations*. We remark that we do not consider the results of their Machine Translation[7] method, since, as emphasised by the authors, it is not informative in terms of cross-lingual properties of AMR (Damonte and Cohen, 2018) because it performs English AMR parsing. We provide details of our model hyperparameters in Appendix C.

Results In Table 2 we show the Smatch[8] score of the models. This metric computes the degree of overlap of two AMR graphs (Cai and Knight, 2013).

We point out the low score of the $\emptyset$-shot models, i.e., XL-AMR$_{\emptyset}^{amr}$ and XL-AMR$_{\emptyset}^{par+}$, which perform lower than AMREAGER, especially in the Chinese language. However, XL-AMR$_{\emptyset}^{par+}$ noticeably improves over XL-AMR$_{\emptyset}^{amr}$, which can be explained by the fact that *seq2seq* requires a large amount of data in order to generalize. This is confirmed by a fine-grained analysis showing lower accuracy of XL-AMR$_{\emptyset}^{amr}$ compared to XL-AMR$_{\emptyset}^{par+}$ in concept identification, which, we recall, is a *seq2seq* module.

Interestingly, the language-specific XL-AMR^{par}, even if trained on less instances, outperforms the $\emptyset$-shot models by a large margin. Moreover, it also surpasses AMREAGER, which is trained on the same sentences from Europarl. The results are further improved when jointly training in multiple languages, i.e., when using the multilingual and bilingual configurations. We attribute this improvement to the ability of a *seq2seq* model to learn better when provided with a larger training set. The domain of the Europarl data is very specific, which does not enable the model to generalize in sentences from other domains. In fact, the XL-AMR^{par+} models significantly improve over the XL-AMR^{par} bilingual and multilingual models. We attribute the higher performances of XL-AMR^{par+} to i) larger training dataset, ii) training on different domains, and iii) better quality of the data (AMR 2.0 data is human annotated).

The XL-AMR^{trans} models perform best: we note that the performances of the language-specific variants outperform those of the multilingual XL-AMR^{trans} models, in contrast to the behaviour of the XL-AMR^{par} models, suggesting that the addition of silver data in other languages is not beneficial. This may be due to the fact that the AMR graphs of translated sentences are the same, thus as a consequence the model does not access extra information. Moreover, the inclusion of translated sentences in other languages slightly harms the performances. This is confirmed by the removal from the training set of the most distant language, in the multilingual (-ZH) model, which in turn achieves around 2 F1 points more compared to the multilingual version including Chinese. This can be further explained by the linguistic differences between Chinese and the other languages, which prevent them from benefiting from the inclusion of Chinese instances in the training set. However, when adding English gold AMR 2.0, i.e., XL-AMR^{trans+}, the model benefits from the better quality of this dataset. In fact, the bilingual version of XL-AMR^{trans+} is the best performing across the board in German, Spanish and Italian, surpassing AMREAGER by at least 14 F1 points and both XL-AMR^{par} and XL-AMR^{par+} by at least 5 F1 points in each language. Interestingly, the best results in Chinese are achieved by the language-specific XL-AMR^{trans} surpassing AMREAGER by 8 F1 points and the $\emptyset$-shot models by more than 17 F1 points. This is once again explained by the linguistic differences of Chinese as compared to the other languages, which render the additional data non-beneficial. Table 3 shows the fine-grained evaluation of AMREAGER and our best performing models for each data creation approach, for which we

[7] It translates the test sentences from the target language to English and parses the translations using an English parser.

[8] github.com/snowblink14/smatch

	AMREAGER				XL-AMR^{par+}				XL-AMR^{trans+}			
Metric	DE	ES	IT	ZH	DE	ES	IT	ZH	DE	ES	IT	ZH
SMATCH	39.1	42.1	43.2	34.6	47.0	53.0	51.4	-	**53.0**	**58.0**	**58.1**	**43.1**
Unlabeled	45.0	46.6	48.5	41.1	52.0	58.3	57.1	-	**57.7**	**63.0**	**63.4**	**48.9**
No WSD	39.2	42.2	42.5	34.7	47.1	53.2	51.5	-	**53.2**	**58.4**	**58.4**	**43.2**
Reentrancies	18.6	27.2	25.7	15.9	33.6	40.1	39.2	-	**39.9**	**46.6**	**46.1**	**34.7**
Concepts	44.9	53.3	52.3	39.9	48.7	58.0	55.6	-	**58.0**	**65.9**	**64.7**	**48.0**
Named Ent.	63.1	65.7	67.7	67.9	63.1	61.6	62.7	-	**66.0**	**66.2**	**70.0**	60.6
Wikification	49.9	44.5	50.6	46.8	61.4	63.8	66.1	-	**60.9**	**63.1**	**67.0**	**54.5**
Negation	**18.6**	19.8	22.3	6.8	8.1	21.5	25.7	-	11.7	**23.4**	**29.2**	**12.8**
SRL	29.4	35.9	34.3	27.2	40.8	48.7	46.7	-	**47.9**	**55.2**	**54.7**	**41.3**

Table 3: Fine-grained F1 scores DE, ES, IT and ZH. Best scores per language are denoted in **bold**.

use the evaluation tools[9] of Damonte et al. (2017). The fine-grained results for the AMREAGER are not reported by Damonte and Cohen (2018), therefore we run the evaluation using their released models.[10] Our best model outperforms AMREAGER in all subtasks except for *Negations* in German and *Named Entities* in Chinese, which are prone to heuristic string matching errors in the pre- and postprocessing procedure of our models. XL-AMR^{trans+} achieves significantly higher performance in *Reentrancies*, *Concepts*, *SRL*, in all the tested languages, compared to AMREAGER, thus demonstrating the effectiveness of our parser and data creation approaches.

In summary, translating the gold standard training data, i.e., GOLDAMR-SILVERTRNS, leads XL-AMR to achieve higher performances than when trained on parallel sentences associated with silver AMR graphs, i.e., PARSENTS-SILVERAMR.

5 Qualitative Analysis

We manually check the predictions of XL-AMR in order to establish the nature of the mistakes based on the Smatch score between the gold and predicted AMR graphs and determine their severity. Then, we observe how XL-AMR handles the translation divergences, i.e., linguistic distinctions that make transfer across languages difficult (Dorr, 1994).

Smatch errors The parser has difficulties with some compounded words in German, e.g., *Uranproduktionsfähigkeit* (*uranium production capability*), *Kernkraftstoffkreislauf* (*nuclear fuel cycle*), for which it fails to break their meaning down to the correct subgraph, e.g., (c / cycle-02 :ARG1 (f /fuel :mod (n / nucleus))), thus predicting a generic node, i.e., (t / thing). This issue can be alleviated using a better preprocessing to split the compounds.

Several cases with low Smatch score are due to inconsistent translations of test set sentences into the target language, even though, we recall, the test set has been manually translated. This could be due to translator choices, but can lead to divergent meaning structures, e.g., *Ich kann verstehen, wie Du Dich fühlst (DE)* (*I can understand how you are feeling*) whose original English sentence from which the AMR graph is projected is *I know what you're feeling*. The gold AMR graph is thus not appropriate for the German sentence, due to the sentence's different meaning. Thus these mistakes are not due to the parser, but to the translations.

An interesting cause of drop in the Smatch arises from the prediction of concepts that are synonyms of the corresponding concepts in the gold graph, e.g., say-01 → state-01, stop-01 → halt-01, best friend → best mate, demand-01 → urge-01, etc. We notice that the predicted concepts (to the left of the arrow) are less specific than the gold concepts, yet somehow preserve the meaning. These examples show that the parser captures a close meaning even when failing to predict the exact concept.

Translation divergences We investigate how XL-AMR deals with the cases where there exist translation divergences, i.e., cases in which source and target language have different syntactic ordering properties (Dorr, 1990), as classified by Dorr (1994) using the following 7 categories: i) *thematic*, ii) *promotional*, iii) *demotional*, iv) *structural*, v) *conflational*, vi) *categorial*, vii) *lexical*.[11]

A *thematic* divergence happens when the argument-predicate structure is different across lan-

[9] github.com/mdtux89/amr-evaluation

[10] github.com/mdtux89/amr-eager-multilingual

[11] In absence of a larger available resource for language divergences, here we make use of some of the pre-classified examples from Dorr (1990, 1994).

guages, e.g., ***I*** *like travelling* where *I* is the subject, in Italian becomes ***Mi*** *piace viaggiare*, and *Mi* is now the object. XL-AMR overcomes this divergence and predicts the correct AMR, (l / like-01 :ARG0 (i / I) :ARG1 (t / travel :ARG0 i)).

Promotional and *demotional* divergences can be merged into the *head switching* macro-category. They arise when a modifier in one language is promoted to a main verb in the other, or vice versa, e.g., *John* ***usually*** *goes home* is *Juan* ***suele*** *ir a casa* (*John is accustomed to go home*) in Spanish. XL-AMR correctly parses the sentence into (g / go-01 :ARG0 (p / person :name (n / Juan)) :ARG4 (h / home) :mod (u / usual)).

A *structural* divergence exists when a verbal object is realized as a noun phrase (NP) in one language and as prepositional phrase (PP) in the other, e.g., *I saw* ***John*** where *John* is NP, is translated as *Vi* ***a Juan*** (*I saw to John*) in Spanish where *a Juan* is PP. This also is not a problem for our parser, which predicts the correct graph, (s / see-01 :ARG0 (i / I) :ARG1 (p / person :name (n / Juan))).

A *conflational* divergence refers to the translation of two or more words in one language into one word in the other. The above errors in German compounded words fall into this category and our model does not handle them properly. However, regarding other languages this problem is not common, e.g., *I* ***fear*** translates into *Io* ***ho paura*** (*I have fear*) in Italian and the parser correctly predicts the AMR graph, (f / fear-01 :ARG0 (i / I)).

A *categorical* divergence arises when the same meaning is expressed by different syntactic categories across languages, e.g., *I* ***agree***, where *agree* is a *verb*, is expressed by a *noun* in Italian and Spanish, *Sono d’**accordo*** and *Estoy de* ***acuerdo***. The parser correctly predicts the same AMR for both languages, (a / agree-01 :ARG0 (i / I)).

A *lexical* divergence arises when a verb in the source language is translated with a different lexical verb, e.g., *John* ***broke*** *into the room*, *Juan* ***forzó*** *la entrada al cuarto*, in which the verb *break* in English is translated with the verb *forzar* (*force*) in Spanish. XL-AMR predicts (f / force-01 :ARG0 (p / person :name (n / Juan)) :ARG2 (e / enter-01 :ARG0 p :ARG1 (r / room))) for the Spanish sentence, which, even though it is correctly parsed, does not overcome the lexical difference of the action, which results in different AMR graphs for the same meaning. This is partially due to the fact that AMR is bounded to lexical forms in English.

In summary, XL-AMR overcomes most of the foregoing structural divergences with the exception of two cases: i) the conflational divergence in German, that is caused by the language’s compound words vocabulary, for the resolution of which a better preprocessing can be beneficial; ii) the lexical divergence that persists despite the parser predicting a valid graph. The latter divergence results in non-parallel structures for parallel meanings, and we believe this might be tackled by integrating a unified ontology for synonyms or related meanings within the AMR formalism, along the line of disjunctive AMR[12] (Banarescu et al., 2013). We leave exploration of this approach open for future work.

6 Conclusion

We explored transfer learning techniques to enable high performance cross-lingual AMR parsing. We created silver data based on annotation projection through parallel sentences and machine translation, on which we trained XL-AMR, a cross-lingual AMR parser that achieves the highest results reported to date on Chinese, German, Italian and Spanish. A qualitative evaluation showed that XL-AMR is able to handle most of the structural divergences among languages. The performance of XL-AMR together with the qualitative analysis suggests that carefully modeling cross-lingual AMR parsing leads to the production of suitable AMR structures across languages. It would therefore be promising to extend this line of our research to exploit larger multilingual semantic resources, in order to further improve the parsing quality. These AMR representations could then be integrated into downstream cross-lingual tasks to investigate their added value.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the ELEXIS project No. 731015 under the European Union’s Horizon 2020 research and innovation programme.

This work was partially supported by the MIUR under the grant “Dipartimenti di eccellenza 2018-2022” of the Department of Computer Science of Sapienza University.

The authors would like to thank Luigi Procopio for the valuable discussions during this work.

[12] amr.isi.edu/damr.1.0.pdf

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China.

Alessio Palmero Aprosio and Giovanni Moretti. 2016. Italy goes to stanford: a collection of corenlp modules for italian. *CoRR*, abs/1609.06204.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.

Edoardo Barba, Luigi Procopio, Niccolò Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. MuLaN: Multilingual label propagatioN for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3837–3844.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 684–695, Marseille, France.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria.

Simone Conia and Roberto Navigli. 2020. Conception: Multilingually-enhanced, human-readable concept vector representations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 121–124.

Marco Damonte and Shay Cohen. 2020. *Abstract Meaning Representation 2.0 - Four Translations LDC2020T07*. Web Download, Philadelphia: Linguistic Data Consortium.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual abstract meaning representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, New Orleans, Louisiana.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. An incremental parser for Abstract Meaning Representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 536–546, Valencia, Spain.

Marco Damonte, Rahul Goel, and Tagyoung Chung. 2019. Practical semantic parsing for spoken language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 16–23, Minneapolis - Minnesota.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China.

Bonnie Dorr. 1990. Solving thematic divergences in machine translation. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 127–134, Pittsburgh, Pennsylvania, USA.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland.

DongLai Ge, Junhui Li, Muhua Zhu, and Shoushan Li. 2019. Modeling source syntax and semantics for neural AMR parsing. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4975–4981.

Jan Hajič, Ondřej Bojar, and Zdeňka Urešová. 2014. Comparing Czech and English AMRs. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 55–64, Dublin, Ireland.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English dependency treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3153–3160, Istanbul, Turkey.

Hardy Hardy and Andreas Vlachos. 2018. Guided neural language generation for abstractive summarization using Abstract Meaning Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 768–773, Brussels, Belgium.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Fuad Issa, Marco Damonte, Shay B. Cohen, Xiaohui Yan, and Yi Chang. 2018. Abstract meaning representation for paraphrase detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 442–452, New Orleans, Louisiana.

Yoon Kim, Yacine Jernite, David A. Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2741–2749.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Paul Kingsbury and Martha Palmer. 2002. From Tree-Bank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200, Sofia, Bulgaria.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yijia Liu, Wanxiang Che, Bo Zheng, Bing Qin, and Ting Liu. 2018. An AMR aligner tuned by transition-based parser. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2430, Brussels, Belgium.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

Chunchuan Lyu and Ivan Titov. 2018. AMR parsing as graph prediction with latent alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 397–407, Melbourne, Australia.

Weimin Lyu, Sheng Huang, Abdul Rafae Khan, Shengqiang Zhang, Weiwei Sun, and Jia Xu. 2019. CUNY-PKU parser at SemEval-2019 task 1: Cross-lingual semantic parsing with UCCA. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 92–96, Minneapolis, Minnesota, USA.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 4585–4592, Reykjavik, Iceland.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli. 2018. Natural language understanding: Instructions for (present and future) use. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5697–5702.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Rik van Noord and Johan Bos. 2017. Neural semantic parsing by character-based translation: Experiments with abstract meaning representations. *CoRR*, abs/1705.09980.

Stephan Oepen and Jan Tore Lønning. 2006. Discriminant-based MRS banking. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Tommaso Pasini. 2020. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-20, Yokohama, Japan*.

Xiaochang Peng, Chuan Wang, Daniel Gildea, and Nianwen Xue. 2017. Addressing the data sparsity issue in neural AMR parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 366–375, Valencia, Spain.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey.

Nima Pourdamghani, Yang Gao, Ulf Hermjakob, and Kevin Knight. 2014. Aligning English strings with abstract meaning representation graphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 425–429, Doha, Qatar.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online.

Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. Biomedical event extraction using Abstract Meaning Representation. In *BioNLP 2017*, pages 126–135, Vancouver, Canada,.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019a. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019b. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451.

Jörg Tiedemann. 2015. Improving the cross-lingual projection of syntactic dependencies. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 191–199, Vilnius, Lithuania.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Kai Von Fintel and Lisa Matthewson. 2008. Universals in semantics. *The linguistic review*, 25(1-2):139–201.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas.

Nianwen Xue, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Urešová, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1765–1772, Reykjavik, Iceland.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. AMR parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 80–94, Florence, Italy.

Sheng Zhang, Xutai Ma, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2018. Cross-lingual decompositional semantic parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1664–1675, Brussels, Belgium.

Huaiyu Zhu, Yunyao Li, and Laura Chiticariu. 2019. Towards universal semantic representation. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 177–181, Florence, Italy.

A Cross-Lingual AMR Pre- and Postprocessing

English AMR parsers throughout the literature rely on several pre- and postprocessing rules. We extend these rules for the cross-lingual AMR parsing task based on several multilingual resources.

Preprocessing This step consists of: i) lemmatization, ii) PoS-tagging, iii) NER, iv) re-categorization of entities and senses and v) removal of wiki links and polarity attributes. As NLP pipelines (steps i-iii) we use Stanford CoreNLP (Manning et al., 2014) for English sentences, the Stanza Toolkit (Qi et al., 2020) for Chinese, German and Spanish sentences, and Tint[13] (Aprosio and Moretti, 2016) for Italian. Re-categorization and anonymization of entities is often used in English AMR parsing to reduce data sparsity (Zhang et al., 2019; Lyu and Titov, 2018; Peng et al., 2017; Konstas et al., 2017). Here we follow Konstas et al. (2017); Zhang et al. (2019) and anonymize entity subgraphs, which are identified by an AMR entity type and the `:name` role. First, the entity subgraphs are mapped with the corresponding text span in the sentence and then the text span is replaced with the anonymized token, i.e., `ENTITY_TYPE_i`. To match the entities in the AMR graphs, which are tied to English, with the corresponding text span in non-English sentences, we first collect all the possible lexicalizations of the entity in the target language using BabelNet 4.0 (Navigli and Ponzetto, 2010), a multilingual semantic network which brings together different resources such as WordNet, Wikipedia, etc., each node of which clusters together the lexicalizations that express the same concept in different languages. Then we search for the possible text spans in the sentence written in the target language. At test time, we anonymize the text spans which have been identified during the training data preprocessing and which are tagged by the NER tagger as entities.

Postprocessing This step consists of restoring i) anonymized subgraphs, ii) wiki links, iii) senses and iv) polarity attributes. The anonymized subgraphs are restored using the anonymized text

[13] Stanza does not provide a NER model for Italian.

spans created during preprocessing. Then wiki links are restored using the DBpedia Spotlight API[14] (Daiber et al., 2013), commonly used in English AMR parsing (van Noord and Bos, 2017; Zhang et al., 2019; Ge et al., 2019). It provides models for multiple languages, except Chinese, for which we use Babelfy (Moro et al., 2014). Since the wiki links identified by DBpedia Spotlight API are language-specific to the text, we further use Wikipedia inter-language links to retrieve the corresponding wiki links for the English entities. We restore senses as the most frequent sense of the predicate in the training data (using -01 if unseen) similar to (Lyu and Titov, 2018; Zhang et al., 2019) and finally restore polarity attributes based on heuristic rules observed on the training data and linguistic rules specific to each language (included in the released code).

B OpusMT Translation Models

For the translation and back-translation steps of GOLDAMR-SILVERTRNS data creation approach, we use the pretrained models[15] from the huggingface transformers library[16] listed in Table 4.

Source	Target	Model
German	English	`Helsinki-NLP/opus-mt-de-en`
Italian	English	`Helsinki-NLP/opus-mt-it-en`
Spanish	English	`Helsinki-NLP/opus-mt-ROMANCE-en`
English	German	`Helsinki-NLP/opus-mt-en-de`
English	Italian	`Helsinki-NLP/opus-mt-en-it`
English	Spanish	`Helsinki-NLP/opus-mt-en-ROMANCE`

Table 4: OpusMT translation models.

C Model Hyperparameters

The input features for all the models include: i) fixed mBERT[17] (Devlin et al., 2019) as contextual embeddings (dim = 768), ii) ConceptNet Numberbatch 9.08[18] (Speer et al., 2017) multilingual static word embeddings (dim = 300) which we set as trainable except in $\emptyset$-shot models, iii) trainable PoS embeddings (dim = 100) where we use the universal PoS-tags set by Petrov et al. (2012), iv) trainable anonymization indicator embeddings (dim = 50), v) trainable character-level embeddings (dim = 100), i.e., CharCNN (Kim et al., 2016).

The encoder and decoder of the node prediction module are composed of 2 layers of 512 and 1024 LSTM units each, respectively. All the models are trained using Adam optimizer (Kingma and Ba, 2015) with learning rate 0.001, for 120 epochs and the best model hyperparameters are chosen on the basis of development set accuracy. The models are trained using 1 GeForce GTX TITAN X GPU, full training takes around 48 hours for models trained in the largest dataset XL-AMR^{trans+} (~84M trainable parameters) and XL-AMR^{par+} (~86M trainable parameters). At prediction time we set the size of beam search to 5.

[14] github.com/dbpedia-spotlight/spotlight-docker.

[15] 6-layer Transformer-based models (Vaswani et al., 2017).

[16] huggingface.co/transformers/model_doc/marian.html

[17] `bert-base-multilingual-cased`: a contextualized embedding for a token is calculated as the average pooling of its subtoken embeddings.

[18] github.com/commonsense/conceptnet-numberbatch