Maximum Bayes Smatch Ensemble Distillation for AMR Parsing

Young-Suk Lee†, Ramón Fernandez Astudillo†, Thanh Lam Hoang‡,
Tahira Naseem†, Radu Florian†, Salim Roukos†
{ysuklee,tanseem,raduf,roukos}@us.ibm.com
ramon.astudillo@ibm.com
t.l.hoang@ie.ibm.com
IBM Research AI†, IBM Research - Ireland‡

Abstract

AMR parsing has experienced an unprecedented increase in performance in the last three years, due to a mixture of effects including architecture improvements and transfer learning. Self-learning techniques have also played a role in pushing performance forward. However, for most recent high performant parsers, the effect of self-learning and silver data generation seems to be fading. In this paper we show that it is possible to overcome this diminishing returns of silver data by combining Smatch-based ensembling techniques with ensemble distillation. In an extensive experimental setup, we push single model English parser performance above 85 Smatch for the first time and return to substantial gains. We also attain a new state-of-the-art for cross-lingual AMR parsing for Chinese, German, Italian and Spanish. Finally we explore the impact of the proposed distillation technique on domain adaptation, and show that it can produce gains rivaling those of human annotated data for QALD-9 and achieve a new state-of-the-art for BioAMR.

1 Introduction

The adoption of the Transformer architecture (Vaswani et al., 2017) for Abstract Meaning Representation (AMR) parsing (Cai and Lam, 2020; Fernandez Astudillo et al., 2020) as well as pretrained language models (Bevilacqua et al., 2021; Zhou et al., 2021b) have enabled an improvement of above 10 points in the last two years, as measured with the standard Smatch (Cai and Knight, 2013) metric. Data augmentation techniques have also shown great success in pushing the state-of-the-art of AMR parsing forward. These include generating silver AMR annotations with a trained parser (Konstas et al., 2017; van Noord and Bos, 2017), with multitask pre-training and fine-tuning (Xu et al., 2020) as well as combining AMR to source text and silver AMR generation (Lee et al., 2020) and stacked pre-training of silver data from different models – from low performance to high performance silver data (Xia et al., 2021).

Despite this remarkable progress, the latest BART-based state-of-the-art parsers, have shown diminishing returns for data augmentation. Indeed both SPRING (Bevilacqua et al., 2021) and Structured-BART (Zhou et al., 2021b) gain a mere 0.5 Smatch from self-learning, compared with over 1 point gains of the previous, less performant, models. Since performance scores are already above where inter annotator agreement (IAA) is assumed to be, i.e. 83 for newswire and 79 for web text reported in (Banerescu et al., 2013), one possible explanation is that we are reaching some unavoidable performance plateau.

In this work we show that this is not the case and that by leveraging the right data augmentation technique one can return to prior gains. The main contributions of this paper are as follows:

- We propose to combine previous techniques such as Smatch-based model ensembling (Barzdins and Gosko, 2016; Hoang et al., 2021) and ensemble distillation (Hinton et al., 2015) of heterogeneous parsers to produce high quality silver data.

- We offer a Bayesian ensemble interpretation of this technique as alternative to views such as Minimum Bayes Risk decoding (Goel and Byrne, 2000) and name the technique Maximum Bayes Smatch Ensemble (MBSE hereafter).

- Applied to English monolingual parsing, MBSE distillation yields a new single system state-of-the-art (SoTA hereafter) on both AMR2.0 (85.4) and AMR3.0 (83.8) test sets. Combined with AMR-to-text data augmentation of Lee et al. (2020) and Structured-mBART in Section 4.3, it yields new SoTA for Chinese (74.2), Italian (75.7) and Spanish (76.9) cross-lingual parsing.

- Applied to domain adaptation, MBSE distillation achieves the performance comparable to adding human annotations of QALD-9 training data and achieves new SoTA on BioAMR test set.

- Ablation studies show that the effectiveness of MBSE distillation lies in the the Smatch-driven selection, not achievable by a random mix of highest accuracy models, under the same training condition.
Figure 1: Data augmentation framework for AMR parsing. Given a labeled example in English \((e_1, g_1)\), we use an AMR-to-text generation system to generate an alternative input text \(\hat{e}_2\) for \(g_1\). Maximum Bayes Smatch decoding produces annotations for unlabeled English sentences. Given a sentence \(e_3\), it ensembles various state-of-the-art off-the-shelf parser outputs and re-ranks them based on their mutual Smatch agreement, to derive the unsupervised annotation \(g_3\). Translating the English sentences to (for example) Spanish, yields the new training samples \((\hat{e}_1, g_1), (\hat{e}_2, g_1), (\hat{e}_3, g_3)\) to train a Spanish cross-lingual parser. We can also use the English pairs \((e_1, g_1), (e_2, g_1), (e_3, g_3)\) to train an English parser. We use Watson Language Translator (WLT) for machine translation.

## 2 Maximum Bayes Smatch Ensemble

Ensemble distillation integrates knowledge of different teacher models into a student model. For sequence to sequence models, e.g. machine translation, it is possible to ensemble models by combining probabilities for words given context at each time step (Kim and Rush, 2016; Freitag et al., 2017). Syntactic and semantic parsers model a distribution over graphs that is harder to integrate across teacher models in an optimal way. For particular cases like dependency parsing, it is possible to ensemble teachers based on the notion of edge attachment (Kuncoro et al., 2016), well related to the usual evaluation metric, Label Attachment Score (LAS). For AMR graphs, however, this is not possible. AMR graphs are quite complex and not explicitly aligned to words. The standard metric for comparing two AMR graphs, Smatch (Cai and Knight, 2013) approximates the NP-Complete problem of aligning nodes across graphs with a hill climbing algorithm. This illustrates the difficulty of achieving consensus across teachers for AMR ensembling.

Prior work ensembling AMR graphs has leveraged Smatch directly or its hill climbing strategy for ensembling. The ensemble in (Barzdins and Gosko, 2016) selects, among a number of candidate AMRs, the one that has the largest average Smatch with respect to all sampled AMRs. The ensemble in (Hoang et al., 2021), modifies the candidate AMRs to increase consensus as measured by coverage. Then it selects from the union of original and modified graphs for the one with highest coverage or largest average Smatch. One possible interpretation of both techniques is that of Minimum Bayes Risk (MBR) decoding, a well established method in Automatic Speech Recognition (ASR) (Goel and Byrne, 2000) and Machine Translation (MT) (Kumar and Byrne, 2004). Assuming that we have a model predicting a graph from an input sentence \(p(g | w)\), normal decoding entails searching among model outputs \(g\) for the one that has the highest likelihood according to the model \(p(g | w)\). MBR searches instead for the model output that minimizes the risk with respect to the distribution of possible human (gold) outputs for a given input

\[
\hat{g} = \arg \min_g \left\{ E_{p(g^h | w)} \left\{ R(g, g^h) \right\} \right\}
\]

where \(p(g^h | w)\) is the distribution of correct human outputs, e.g. given by multiple annotators, and \(R\) is a risk function that measures how severe deviations from \(g^h\) are, in this case minus Smatch. Since in practice \(p(g^h | w)\) is not available, MBR takes often the strong assumption of replacing \(p(g^h | w)\) by the model distribution itself \(p(g | w)\).

Here we suggest another Bayesian interpretation, that requires less strong assumptions than MBR, a Bayesian model ensemble (Wilson and Izmailov, 2020). Indeed techniques above can be seen as solving

\[
\hat{g} = \arg \max_{g \in G} \left\{ E_{p(M | D)} \left\{ \text{Smatch}(g, \tilde{g}_M) \right\} \right\}
\]

where \(p(M | D)\) is the distribution of models \(M\) given training data \(D\), approximated by a sample average of models of different architectures or different random seeds, and

\[
\tilde{g}_M = \text{post} \left( \arg \max_{y} \left\{ \prod_{t=1}^{y} p_M(y_t | y_{<t}, w) \right\} \right)
\]

is the output of a conventional decoding process for each parser prediction distribution \(p_M\), including post-processing \(\text{post}()\). This process differs across models indexed by \(M\), for example \(y\) can be transition actions or linearized graphs and \(\text{post}()\) the state-machine or linearized graph post-processing\(^1\). \(\tilde{g}\) is the space of

\(^1\)We consider only auto-regressive models in this work but this approach could also encompass e.g. graph-based parsers.
candidate graphs, which in Barzdins and Gosko (2016) are the AMRs resulting from decoding each sample from \(p(M | D)\) and in Hoang et al. (2021) are those same graphs plus the modified pivot graphs i.e. twice that number of candidates. There is in principle no restriction on how to build the set \(\mathcal{G}\). Decoding a graph \(g \in \mathcal{G}\) means here selecting the member of that set maximizing the expected Smatch and is different from each parser’s decoding process.

If we replace \(\text{Smatch}(\cdot)\) by an indicator function on the decoding outputs \(1_{\tilde{g} = g}\), we recover majority voting of AMR graphs. Since the space of graphs is exponentially large on the input size, this would be too sparse to attain meaningful vote counts. The propagation of the uncertainty in \(p(M | D)\) through the decoding and \(\text{Smatch}(\cdot)\) transformations both solves the sparsity problem, and allows optimization on a space that is better related to the target metric. The method will be henceforth described here as Maximum Bayes Smatch Ensemble distillation (MBSE distillation). In what follows, we will consider two versions for distillation, one utilizing the Smatch version of Hoang et al. (2021), and another utilizing a greedy version of Barzdins and Gosko (2016) where we select the two highest Smatch AMRs and from that pair, keep the graph with the highest Smatch with respect to the remaining graphs, henceforth termed ‘greedy-select’. The greedy-select algorithm is given in Algorithm 1 and performs similarly to an implementation of Barzdins and Gosko (2016).

Algorithm 1: Greedy-Select MBSE Algorithm

**Input:** AMR\(_1\)…AMR\(_n\) parses from \(n\) AMR parsing models, where \(n \geq 3\)

**Output:** One-best AMR parse

1: Let bestAMR = NULL
2: for \(\forall i, j\) in \(1 \leq i, j \leq n\) and \(i \neq j\) do
3: Compute sentence Smatch score \(\text{smatch}(\text{AMR}_i, \text{AMR}_j)\), total \(n(n - 1)/2\) scores.
4: Pick the highest \(\text{smatch}(\text{AMR}_i, \text{AMR}_j)\).
5: for Each AMR\(_a\), where \(a = i\) or \(a = j\) do
6: Pick the highest \(\text{smatch}(\text{AMR}_a, \text{AMR}_b)\).
7: if \(a = i\) then
8: bestAMR = AMR\(_i\)
9: else
10: bestAMR = AMR\(_j\)
11: end if
12: end for
13: end for
14: return bestAMR

3 Silver Training Strategy

We now describe the AMR silver training strategy proposed in this work. This strategy creates high quality English and cross-lingual AMR annotations for unlabeled data with MBSE and alternative input sentences of gold AMRs via AMR-to-text generation.

As depicted in Fig. 1, we start with 1) a set of gold-labeled (English sentence, AMR) pairs, 2) a set of unlabeled English sentences and 3) pre-trained English-to-foreign language Machine Translation systems to obtain foreign language input sentences corresponding to English AMRs. Assuming \(N\) off-the-shelf AMR parsers, we train each of the \(N\) parsers using the gold data with their respective training procedure. More than one random seed may be trained for some parsers, leading to more than \(N\) AMR parses for each input sentence.

After the parsers have been trained, we use them to parse the unlabeled English text as in Konstas et al. (2017). Interpreting the set of trained models as samples of the model distribution, we apply the MBSE distillation methods described in Section 2. In the case of greedy-select, we compute sentence-level Smatch scores across all outputs. For each sentence, we select the parse that maximally agrees with the other parses as the one best output according to Algorithm 1. This yields one MBSE-distilled silver AMR annotation for every unlabeled sentence.

For English parsers, the MBSE-distilled AMR annotations are added to human-annotated gold treebanks for enhanced model training. For cross-lingual parsers, we use MT systems to translate all English input sentences corresponding to both human annotated and MBSE-distilled silver data to the target foreign languages and train respective cross-lingual parsers with pairs of (Foreign language input sentences, AMR graphs in English), following (Damonte and Cohen, 2018).

In addition, following Lee et al. (2020), we train an AMR-to-text model based on pre-trained Transformers (Mager et al., 2020; Ribeiro et al., 2020; Bevilacqua et al., 2021) from human annotated treebanks for English, and use it to generate additional sentences for every human-annotated sentence. We filter out the generated texts if they are too similar or too dissimilar to the original input texts, as measured by the BLEU score of (Papineni et al., 2002), keeping only those for which BLEU > 0.1 and BLEU < 0.9. The data augmented with AMR-to-text generation\(^2\) is used for cross-lingual AMR parser training along with the MBSE-distilled silver training data.

4 Experimental Setup

4.1 Corpus Statistics

Table 1 details the main corpora considered for the standard benchmark experiments on AMR2.0 and AMR3.0 test sets (top) and out-of-domain data used for domain adaptation experiments (bottom). Silver

\[^{2}\]We use SPRING of Bevilacqua et al. (2021) and the pre-trained model from https://github.com/SapienzaNLP/spring.
in the table indicates the unlabeled data used for silver training data acquisition. For domain adaptation, SQuAD2.0-questions are used for QALD-9 data and PubMed and BioNLP-2011 for BioAMR data.

| Dataset                  | Split   | Sents | Tokens |
|--------------------------|---------|-------|--------|
| AMR2.0 (LDC2017T10)     | Train   | 36,521| 653K   |
|                          | Test    | 1,371 | 30K    |
|                          | Dev.    | 1,368 | 29K    |
| AMR3.0 (LDC2020T02)     | Train   | 35,635| 1M     |
|                          | Test    | 1,898 | 39K    |
|                          | Dev.    | 1,722 | 37K    |
| PropBank                 | Silver  | 20K   | 386K   |
| SQuAD2.0-contexts       | Silver  | 70K   | 2M     |

Table 1: Corpus statistics for the standard benchmark experiments on AMR2.0 and AMR3.0 test sets (top) and domain adaptation experiments (bottom). Silver indicates the unlabeled data for silver training data acquisition.

4.2 Parsing Models

We use 3 off-the-shelf AMR parsers to parse unannotated raw texts for ensemble distillation. We train the parsers following their standard configurations.

APT (Zhou et al., 2021a)\(^3\) is a transition-based parser that combines hard attention over sentences with a target side action pointer mechanism to decouple source tokens from node representations and address alignments. APT includes 6 layers and 4 attention heads for both the Transformer encoder and decoder, with model size 256 and feed-forward size 512. Cross-attention of all decoder layers is used for action-source alignment.

SPRING Bevilacqua et al. (2021)\(^4\) fine-tunes an AMR parser on pre-trained BART (Lewis et al., 2019) large model, a sequence-to-sequence model trained by corrupting text with an arbitrary noise function and learning to reconstruct the original text. It focuses on predicting linearized AMR graphs using BART and touts its end-to-end system simplicity, avoiding complex pipelines or any heuristics.

Structured-BART Zhou et al. (2021b)\(^5\) combines APT with the pre-trained BART. It focuses on modeling the transition-based parser state within the pre-trained BART architecture and addresses the under-pinnings of AMR parsing such as the graph well-formedness guarantee or built-in graph-sentence alignments. It is the previous SoTA on the standard AMR2.0 and AMR3.0 test sets. We use Structured-BART as the main parser architecture for ensemble distillation.

4.3 Structured-mBART

For cross-lingual AMR experiments, we adapt Structured-BART by replacing the pretrained BART with mBART of (Liu et al., 2020), which we call Structured-mBART.\(^6\) Structured-mBART diverges from Structured-BART mainly in input processing and vocabulary:

- For task vocabulary, Structured-mBART includes $\approx$250K sentencepiece tokens of (Kudo, 2018) including 25 language tags, e.g. en, xx, es, de, de, DE, whereas Structured-BART includes $\approx$50K BPE tokens of (Sennrich et al., 2016).
- For Structured-mBART, we append the source language tag to the end of each input sentence without specifying the target language tag.
- For Structured-mBART, we decrease the learning rate to 3e-5, compared with 1e-4 of Structured-BART, and move the layer normalization to the beginning of each transformer block, as opposed to the end in Structured-BART.

We obtain contextualized embeddings from the pre-trained mBART for multilingual input sentence representations. For the target action sequences, we map the sentencepiece\(^7\) tokens to the corresponding target token, by averaging all values from the sentencepiece tokens corresponding to the target token. For German, Italian and Spanish input texts, we apply the tokenization from JAMR parser\(^8\) before sentencepiece tokenization. For Chinese, we use only the sentencepiece tokenizer for input tokenization.

5 Results

To explore the effect of the proposed ensemble distillation and training strategy, we consider an extensive experimental setup including standard English benchmarks (Section 5.1), cross-lingual benchmarks (Section 5.2) and out of domain data sets (Section 5.3).\(^9\)

We first provide the performance evaluation of each MBSE technique in Table 2 to demonstrate the effectiveness of the ensemble techniques by themselves. We test the algorithm on the standard test data sets from AMR2.0 and AMR3.0 and three out-of-domain data sets, QALD-9, Little Prince (LP) and BioAMR specified in Table 1. We consider here only standard English

\(^6\) We will open-source Structured-mBART.
\(^7\) https://github.com/google/sentencepiece
\(^8\) https://github.com/JFLANigan/jamr
\(^9\) We applied the technique to APT as well, observing similar performance gains by MBSE distillation for all three applications.
AMR parsing. As expected, ensembling techniques largely outperform individual model performances. Greedy-select, similar to Barzdins and Gosko (2016), out-performs Graphene for LP but under-performs for BioAMR. The Smatch version of Graphene outperforms or matches the other two approaches in all data sets.

### 5.1 English AMR Parsing

We consider the standard AMR2.0 (LDC2017T10) and AMR3.0 (LDC2020T02) corpora as gold data. As unlabeled data, we use the sentence examples for every sense in Propbank (LDC2004T14) and the context portion of the SQuAD-2.0 corpus. SQuAD-2.0-contexts originally include 92K sentences, from which we filter out sentences with bad utf8 encoding (7,305) and those for which the APT model produces ill-formed disconnected graphs (14,869).

The results are shown in Table 3. The lower part of the table compares the performances of Structured-BART in various silver data augmentation setups including our proposed MBSE distillation. Note that greedy-select MBSE distillation improves 1.0 Smatch point on AMR2.0 (84.2 vs. 85.2) and 1.5 Smatch point on AMR3.0 (82.0 vs. 83.5) over the Structured-BART baselines and 0.5 against self-trained silver. To isolate the effect of ensembling, we also provide two additional baselines: 1) silver obtained from SPRING, which can be expected to have complementary information to self-trained silver, and 2) an equal mixture of SPRING and Structured-BART (random 50:50), which tests if the Smatch selection strategy bears any effect. MBSE distillation outperforms these two baselines by between 0.2 and 0.5 depending on the scenario, proving that the use of Smatch selection has a clear positive effect in performance and setting a new state-of-the-art for single system of 85.2 for AMR2.0 and 83.5 for AMR3.0. This improvement is also complementary to the one obtainable with conventional ensemble decoding, that pushes the numbers to 85.4 and 83.8 respectively.

| Models                                   | AMR2.0 | AMR3.0 | QALD-9 | LP    | BioAMR |
|------------------------------------------|--------|--------|--------|-------|--------|
| APT (Zhou et al., 2021a)                | 83.0   | 81.1   | 83.7   | 79.0  | 55.2   |
| Structured-BART (Zhou et al., 2021b)    | 84.6   | 83.1   | 87.7   | 81.0  | 62.4   |
| SPRING1 (Bevilacqua et al., 2021)       | 84.2   | 83.2   | 87.7   | 81.3  | 61.6   |
| SPRING2 (Bevilacqua et al., 2021)       | 83.8   | 82.9   | 86.4   | 81.0  | 60.5   |
| Graphene ensemble (Hoang et al., 2021)  | 85.6   | 84.5   | 88.7   | 81.6  | 64.7   |
| Graphene ensemble, Smatch update (Hoang et al., 2021) greedy select ensemble | 86.1 | 84.8 | 88.9 | 82.3 | 65.0 |
|                                           | 85.8   | 84.6   | 89.0   | 82.4  | 63.3   |

Table 2: MBSE (Maximum Bayes Smatch Ensemble) results (lower table) in Smatch over 4 English AMR parsing models (upper table) detailed in Section 4. SPRING1 and SPRING2 indicate 2 SPRING models from 2 random seeds. Highest scores are boldfaced.

34,156 parses from Structured-BART, 25,129 from SPRING1, 17,866 from SPRING2, and 10,235 from APT. MBSE greedy-select silver data for AMR3.0 include 33,200 parses from Structured-BART, 29,407 from SPRING1, 17,830 from SPRING2 and 6,949 from APT. Note that the number of unlabeled data we are using for MBSE silver training is under 90K sentences, which is significantly smaller than those used by previous work, e.g. 1.8M of Xia et al. (2021), 14M of Xu et al. (2020) and 200K of Bevilacqua et al. (2021).

### 5.2 Cross-lingual AMR Parsing

For cross-lingual AMR parsing, we consider the well known cross-lingual extension of AMR2.0 (Damonte and Cohen, 2018). This includes Machine Translation of the training data for German (DE), Spanish (ES), Italian (IT) and Chinese (ZH) and human translated test sets.

Our cross-lingual parsers are trained with Structured-mBART in Section 4.3 and all input sentences of the English training data, including human annotated treebanks and the MBSE silver data, are machine translated into the target languages. Table 4 shows the results on the human translated AMR2.0 test set following standard practices. We provide results for recently published cross-lingual AMR parsers and different silver training versions of Structured-mBART. As reported in (Uhlig et al., 2021), translate-and-parse baselines outperform conventional cross-lingual parsers. Although they require an external MT system and loose access to token to node alignments, they provide a useful comparison. We thus include here translate-and-parse baselines from the combination of WLT and English Structured-BART.

Structured-mBART with MBSE distillation improves the Smatch score by 2.1 to 2.6 over the Structured-mBART baselines, out-performing very strong previous SoTA from (Cai et al., 2021a) on Chinese and Spanish and tied on Italian. Increasing the input sentence diversity via AMR-to-text generation and ensemble decoding further improve the system performances, attaining new cross-lingual SoTA.

---

10SQuAD-2.0 from https://rajpurkar.github.io/SQuAD-explorer/

11We use WLT from https://www.ibm.com/cloud/watson-language-translator for machine translation.
on all four languages. Note that the translate-and-parse pipeline ‘WLT+Structured-BART+MBSE greedy-select silver’ out-performs the highest performing cross-lingual system for all languages except for Spanish.

### 5.3 Domain Adaptation

We use BioAMR targeting the medical domain considered standard references for domain adaptation. We use the AMR2.0 version of BioAMR data as this has clearly defined partitions and was used in recent high performance results in Bevilacqua et al. (2021).\(^\text{12}\) We also include the QALD-9-AMR corpus,\(^\text{13}\) containing human annotated gold AMRs in AMR3.0 style from the QALD-9 corpus.\(^\text{14}\) QALD-9 (Usbeck et al., 2018) is a corpus of natural language questions for executable semantic parsing, used in (Kapanipathi et al., 2021), among other works.

BioAMR is distinct from AMR2.0 treebank data mostly in terms of vocabulary and named entities. QALD-9 differs from AMR3.0 mostly in sentence types rather than vocabulary or named entities. There are much more frequent questions or imperative sentences, as opposed to mostly declarative sentences in AMR3.0.

---

\(^{12}\)amr.isi.edu/download/2016-03-14/amr-release-training-bio.txt, amr-release-dev-bio.txt, amr-release-test-bio.txt

\(^{13}\)We have manually annotated QALD-9 training and test data sets in house, which we will open source.

\(^{14}\)https://github.com/ag-sc/QALD

Table 5 shows the domain adaptation experimental results. Results for SPRING are taken from Bevilacqua et al. (2021). Since BioAMR is annotated in AMR2.0 style and QALD-9-AMR in AMR3.0 style, we use the corresponding Structured-BART models as indicated in the table. For each test set, we report the results in three different training conditions, all of which include either AMR2.0 or AMR3.0 treebank in the training data: (1) use only greedy-select MBSE silver data without any human-annotated in-domain training sentences, indicated by greed. general (silver), greed. domain (silver) and greed. all silver, (2) use human-annotated gold in-domain sentences, indicated by domain gold sentences, (3) use both greedy-select MBSE silver data and human annotated gold in-domain sentences.

| Models | AMR2.0 | AMR3.0 |
|--------|--------|--------|
| Naseem et al. (2019) | 75.5 | - |
| Zhang et al. (2019a) | 76.3±0.1 | - |
| Zhang et al. (2019b) | 77.0±0.1 | - |
| Cai and Lam (2020) | 80.2 | - |
| Fernandez Astudillo et al. (2020) | 80.2±0.0 | - |
| Lyu et al. (2020) | - | 75.8 |
| Lee et al. (2020) (85K silver AMR) | 81.3±0.0 | - |
| Xu et al. (2020) (14M silver AMR) | 81.4 | - |
| Bevilacqua et al. (2021) (SPRING, 200K silver AMR) | 84.5 | 83.0 |
| Zhou et al. (2021a) (APT, 70K silver AMR on AMR2.0) | 82.6±0.1 | 80.3 |
| Xia et al. (2021) (1.8M silver AMR) | 84.2 | - |

| Models | Sep-voc | Joint-voc | Sep-voc | Joint-voc |
|--------|---------|-----------|---------|-----------|
| Structured-BART-baseline | 84.0±0.1 | 84.2±0.1 | 82.3±0.0 | 82.0±0.0 |
| + self-trained silver | - | 84.7±0.1 | 82.7±0.1 | 82.6±0.0 |
| + self-trained silver + ensemble decoding | - | **84.9** | **83.1** | - |
| + SPRING silver | 84.8±0.1 | 84.8±0.0 | 83.0±0.0 | 83.2±0.1 |
| + SPRING + self-trained silver (random 50:50) | 84.8±0.1 | 84.7±0.0 | 83.0±0.0 | 83.2±0.1 |
| + MBSE greedy-select silver | 85.0±0.0 | 85.2±0.1 | 83.4±0.0 | 83.5±0.0 |
| + MBSE greedy-select silver + ensemble decoding | - | **85.4** | - | **83.8** |

Table 3: Smatch scores on AMR2.0 and AMR3.0 test data. Upper rows display the performances of recent published works. Current SoT, Structured-BART is displayed in middle rows in **boldface** and separate results for self-trained silver and ensemble decoding (Zhou et al., 2021b) are also provided. Lower rows show Structured-BART performances with various silver data augmentations and results for separate vocabulary (sep-voc) and joint vocabulary (joint-voc). All methods use the same 90K unlabeled sentences from Table 1 for silver data acquisition described in Section 3. The numbers prefixed by ± indicate the sample standard deviation of Smatch scores across 3 different random seeds.
| Models                                      | LM  | DE  | ES  | IT  | ZH  |
|--------------------------------------------|-----|-----|-----|-----|-----|
| Translate and Parse Pipelines              |     |     |     |     |     |
| Uhrig et al. (2021)                        | 67.6| 72.3| 70.7| 59.1|     |
| WLT+Structured-BART                        | 73.2| 75.9| 75.7| 63.2|     |
| WLT+Structured-BART+MBSE greedy-sel. silver| 73.9| 76.5| 76.1| 63.7|     |
| Cross-lingual Parsers                      |     |     |     |     |     |
| Biloshimi et al. (2020)                    | 53.0| 58.0| 58.1| 43.1|     |
| Sheth et al. (2021) (85K silver AMR)       | 62.7| 67.9| 67.4| –   |     |
| Procopio et al. (2021) (5M parallel corpus)| 69.8| 72.4| 72.3| 58.0|     |
| Cai et al. (2021b)                         | 64.0| 67.3| 65.4| 53.7|     |
| Xu et al. (2021)                           | 70.5| 71.8| 70.8| –   |     |
| Cai et al. (2021a) (320K silver AMR)       | 73.1| 75.9| 75.4| 61.9|     |

Table 4: Cross-lingual Parser Smatch scores on AMR2.0 human translated test sets. mBART\textsubscript{mt} of Procopio et al. (2021) indicates the mBART model fine-tuned on both semantic parsing tasks and the MT data. mBART\textsubscript{mt} of Cai et al. (2021a) indicates an NMT model by (Tang et al., 2020), trained from mBART covering 50 languages. ‘MBSE greedy-sel. silver’ denotes MBSE greedy-select silver data. ‘ens. dec.’ denotes ensemble decoding.

| Models                                      | LM  | DE  | ES  | IT  | ZH  |
|--------------------------------------------|-----|-----|-----|-----|-----|
| Translate and Parse Pipelines              |     |     |     |     |     |
| Uhrig et al. (2021)                        | 69.9±0.0| 74.4±0.3| 73.3±0.2| 59.9±0.0|     |
| WLT+Structured-BART                        | 72.5±0.1| 76.5±0.2| 75.4±0.0| 62.2±0.1|     |
| WLT+Structured-BART+MBSE greedy-sel. silver| 72.9±0.1| 76.6±0.0| 75.6±0.0| 62.3±0.0|     |
| WLT+Structured-BART+MBSE greedy-sel. silver +AMR2Text + ens. dec. | 73.2 | 76.9 | 75.7 | 62.7 |     |

Table 5: Smatch scores for domain adaptation on BioAMR and QALD-9 data sets with varying sizes of human annotated gold data and silver training data for MBSE distillation. All models are trained with Structured-BART.

There have been numerous works applying ensemble/knowledge distillation (Hinton et al., 2015) to machine translation (Kim and Rush, 2016; Freitag et al., 2017; Nguyen et al., 2020; Wang et al., 2020, 2021), dependency parsing (Kuncoro et al., 2016) and question answering (Mun et al., 2018; Ze et al., 2020; You et al., 2021; Chen et al., 2012). The recent (Xia et al., 2021) can also be seen as distilling knowledge from multiple AMR parsers in stages. Regarding ensembling techniques, parameter averaging of checkpoints, as in (Vaswani et al., 2017), or ensemble decoding of models from different seeds (Sutskever et al., 2014;
Vinyals et al., 2015; Zhou et al., 2021a,b) are well established. Regarding ensembling AMR graphs, Barzdins and Gosko (2016) propose choosing the AMR with highest average sentence Smatch to all other AMRs. Hoang et al. (2021) proposed a more complex technique of building new AMRs by exploiting Smatch’s hill-climbing algorithm.

Our work brings together ensemble distillation and Smatch-based ensembling and shows that it can provide substantial gains over the standard self-training. It applies a technique close to Barzdins and Gosko (2016) and offers a unified interpretation of Barzdins and Gosko (2016); Hoang et al. (2021) as a form of Bayesian ensemble in Smatch space. Compared to ensembling techniques outside of AMR, merging at graph level has the advantage of enabling ensembling of heterogeneous parsers, such as transition, graph-based or linearization based models. In general, our work provides evidence that ensemble distillation of heterogeneous models leads to significant gains in AMR parsing when carried out with proper techniques such as MBSE.

Damonte and Cohen (2018) show that it may be possible to use the original AMR annotations devised for English as representations of equivalent sentences in other languages. This line of research has been widely adopted in recent years. Damonte and Cohen (2018); Sheth et al. (2021) propose annotation projection of English AMR graphs to target languages to train cross-lingual parsers, using either statistical or contextualized word alignments. Blloshmi et al. (2020) show that one may not need alignment-based parsers for cross-lingual AMR, and model concept identification as a seq2seq problem. Procopio et al. (2021) reframe semantic parsing as multilingual machine translation (MNMT) and propose a seq2seq architecture fine-tuned on pretrained-mBART with an MNMT objective. Cai et al. (2021b) propose to use bilingual input to enable a model to predict more accurate AMR concepts. Xu et al. (2021) propose a cross-lingual pretraining approach via multitask learning for AMR parsing leveraging both AMR parsing, AMR2Text and machine translation as well as a word-level Kullback-Leibler distillation. Cai et al. (2021a) propose to use noisy knowledge distillation for multilingual AMR parsing, attaining the previous SoTA on Chinese, German, Italian and Spanish cross-lingual parsing. Uhrig et al. (2021) show that a simple pipeline approach without cross-lingual parsing, which translates the target language sentences to English and parse the translated sentences with an English AMR parser, is highly competitive. Compared to these, we introduce a new cross-lingual architecture Structured-mBART adapted from Structured-BART of Zhou et al. (2021b) and attain a new state-of-the-art in Chinese, German, Italian and Spanish cross-lingual parsing by applying the proposed MBSE distillation.

There has been very little work on domain adaptation for AMR parsing. For semantic parsing in general, (Fan et al., 2017) show that a multi-task setup helps transfer learning from an auxiliary task with large labeled data to a target task with smaller labeled data in semantic parsing. (Kannardi et al., 2019) propose a sequential transfer learning method as a domain adaptation method. (Li et al., 2020) proposes a two-stage semantic parsing model for domain adaptation, where the coarse step transfers the domain general structural patterns and the fine step focuses on the difference between domains. Here we subsume domain adaptation for AMR parsing under data augmentation with MBSE distillation where the only difference lies in the unlabeled data. The unlabeled data is drawn from the target domain for the purpose of domain adaptation rather than those similar to the source training data for data augmentation in general.

7 Conclusion
In this paper we proposed a technique called Maximum Bayes Smatch Ensemble (MBSE) distillation, which combines Smatch-based model ensembling Barzdins and Gosko (2016); Hoang et al. (2021) and ensemble distillation Hinton et al. (2015) of heterogeneous parsers, to produce high quality silver data.

Applied to English monolingual parsing, MBSE distillation yields a new single system SoTA on both AMR2.0 (85.4) and AMR3.0 (83.8) test sets. Combined with AMR-to-text data augmentation and a Structured-mBART, it yields new SoTA for Chinese (62.7), German (73.2), Italian (75.7) and Spanish (76.9) cross-lingual AMR parsing.

Applied to domain adaptation of QALD-9, MBSE distillation achieves the performance comparable to human annotations of the in-domain training data. Applied to domain adaptation of BioAMR, MBSE distillation achieves new SoTA with 80.8 Smatch on the BioAMR test set.
References

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. Association for Computational Linguistics.

Guntis Barzdins and Didzis Gosko. 2016. Riga at senseval-2016 task 8: Impact of smatch extensions and character-level neural machine translation on amr parsing accuracy. In Proceedings of SemEval-2016, pages 1143–1147.

Michele Bevilacqua, Rezhina Biloshti, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In AAAI Technical Track on Speech and Natural Language Processing I, Vol. 35 No. 14, pages 12564–12573.

Rezhina Biloshti, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2487–2500, Online. Association for Computational Linguistics.

Deng Cai and Wai Lam. 2020. Amr parsing via graph sequence iterative inference. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1290–1301.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2021a. Multilingual amr parsing with noisy knowledge distillation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2778–2789.

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. Association for Computational Linguistics.

Yitao Cai, Zhe Lin, and Xiaojun Wan. 2021b. Making better use of bilingual information for cross-lingual amr parsing. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1537–1547.

Cen Chen, Chengyu Wang, Minghui Qiu, Dehong Gao, Linbo Jin, and Wang Li. 2012. Cross-domain knowledge distillation for retrieval-based question answering systems. In Proceedings of the World Wide Web Conference 2021.

Marco Damonte and Shy 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. Association for Computational Linguistics.

Xing Fan, Emilio Monti, Lambert Mathias, and Markus Dreyer. 2017. Transfer learning for neural semantic parsing. In Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 48–56.

Ramón Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodgett, and Radu Florian. 2020. Transition-based parsing with stack-transformers. In Findings of the EMNLP2020 (to appear).

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. In arXiv:1702.01802.

Vaibhava Goel and William J. Byrne. 2000. Minimum bayes-risk automatic speech recognition. Computer Speech & Language, 14(2):115–135.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Thanh Lam Hoang, Gabriele Picco, Yufang Hou, Young-Suk Lee, Lam M. Nguyen, Dzung T. Phan, Vanessa López, and Ramon Fernandez Astudillo. 2021. Ensemble graph prediction. OpenReview.

Alvin Kannardi, Gabriela Ferraro, and Qing Wang. 2019. Domain adaptation for low-resource neural semantic parsing. In In Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, pages 87–93.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, María Chang, Cristina Cornelio, Saswati Datta, Achille Fokoue, Dinesh Garg, Alíflo Gliozzo, Saíram Gurajada, Hima Karaman, Naweed Khan, Dinesh Khandelwal, Young-Suk Lee, Yun-Nao Li, Francois Luus, Ndihuwu Makondo, Nandana Mihindukulasooriya, Tahira Naseem, Sumit Neelam, Lucian Popa, Revanth Gangi Reddy, Ryan Riegel, Gaetano Rossiello, Udit Sharma, G P Shrivatsa Bhargav, and Mo Yu. 2021. Leveraging Abstract Meaning Representation for knowledge base question answering. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3884–3894, Online. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1317–1327.

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. Association for Computational Linguistics.
Takahiro Naseem, Abhishek Shah, Hui Wan, Radu Florian, Salim Roukos, and Miguel Ballesteros. 2019. Rewarding smatch: Transition-based amr parsing with reinforcement learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4586–4592.

Xuan-Phi Nguyen, Shaﬁq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversiﬁcation: A simple strategy for neural machine translation. In 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311–318.

Luigi Procopio, Rocco Tripodi, and Roberto Navigli. 2021. Sgl: Speaking the graph languages of semantic parsing via multilingual translation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 325–337.

Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. arXiv preprint arXiv:2007.08426.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725.

Janaki Sheth, Young-Suk Lee, Ramon Fernandez Astudillo, Tahira Naseem, Radu Florian, and Salim Roukos. 2021. Bootstrapping multilingual amr with contextual word alignments. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 394–404.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of NIPS.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and ﬁnetuning. arXiv preprint arXiv:2008.00401.

Sarah Uhrig, Yoalli Rezepka Garca, Juri Opitz, and Anette Frank. 2021. Translate, then parse! a strong baseline for cross-lingual amr parsing. In Proceedings of the 17th International Conference onParsing Technologies (IWPT 2021), pages 58–64.

Ricardo Usbeck, Ria Hari Gusmita, Muhamad Saleem, and Axel-Cyrille Ngonga Ngomo. 2018. 9th challenge on question answering over linked data (qald-9). In Joint proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLiWoD4, Vol-2241, pages 58–64.
Rik van Noord and Johan Bos. 2017. Neural semantic parsing by character-based translation: Experiments with abstract meaning representations. *arXiv preprint arXiv:1705.09980*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Proceedings of NIPS*.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6456–6466.

Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, Chengxiang Zhai, and Tie-Yan Liu. 2020. Transductive ensemble learning for neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 6291–6298.

Andrew Gordon Wilson and Pavel Izmailov. 2020. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*.

Qingrong Xia, Zhenghua Li, Rui Wang, and Min Zhang. 2021. Stacked amr parsing with silver data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4729–4738.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2020. Improving amr parsing with sequence-to-sequence pre-training. *arXiv preprint arXiv:2010.01771*.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2021. Xlpt-amr: Cross-lingual pre-training via multi-task learning for zero-shot amr parsing and text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 896–907.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021. Knowledge distillation for improved accuracy in spoken question answering. In *ICASSP 2021*.

Yang Ze, Linjun Shou, Ming Gong, and Wutao Lin. 2020. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *The Thirteenth ACM International Conference on Web Search and Data Mining*.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019a. Amr parsing as sequence-to-graph transduction. *arXiv preprint arXiv:2010.01771*.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019b. Broad-coverage semantic parsing as transduction. 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 3784–3796.

Jiawei Zhou, Tahira Naseem, Ramon Fernandez Astudillo, and Radu Florian. 2021a. Amr parsing with action-pointer transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5585–5598.

Jiawei Zhou, Tahira Naseem, Ramon Fernandez Astudillo, Young-Suk Lee, Radu Florian, and Salim Roukos. 2021b. Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based amr parsing. In *OpenReview*. 
This figure "decisiontree.png" is available in "png" format from:

http://arxiv.org/ps/2112.07790v1
This figure "distribution2.png" is available in "png" format from:

http://arxiv.org/ps/2112.07790v1
This figure "distribution3.png" is available in "png" format from:

http://arxiv.org/ps/2112.07790v1