Semantically Inspired AMR Alignment for the Portuguese Language

Rafael Torres Anchiêta and **Thiago Alexandre Salgueiro Pardo**

Interinstitutional Center for Computational Linguistics (NILC)

Institute of Mathematics and Computer Science (ICMC)

University of São Paulo (USP), São Carlos, SP, Brazil

`rta@ifpi.edu.br, taspardo@icmc.usp.br`

Abstract

Abstract Meaning Representation (AMR) is a graph-based semantic formalism where the nodes are concepts and edges are relations among them. Most of AMR parsing methods require alignment between the nodes of the graph and the words of the sentence. However, this alignment is not provided by manual annotations and available automatic aligners focus only on the English language, not performing well for other languages. Aiming to fulfill this gap, we developed an alignment method for the Portuguese language based on a more semantically matched word-concept pair. We performed both intrinsic and extrinsic evaluations and showed that our alignment approach outperforms the alignment strategies developed for English, improving AMR parsers, and achieving competitive results with a parser designed for the Portuguese language.

1 Introduction

According to Banarescu et al. (2013), Abstract Meaning Representation (AMR) is a semantic meaning representation, which may be encoded as a rooted Direct Acyclic Graph (DAG) where the nodes are concepts, and the edges are relations among them. This representation explicitly details semantics information, as depicted in Figure 1. In this figure, the `live-01` node is the root of the graph and `city` node introduces a named entity. Moreover, `:ARGx` relations are predicates from the PropBank lexicon (Kingsbury and Palmer, 2002), which encode semantic information according to each PropBank sense.

To parse a text into an AMR graph, most of the AMR parsers require alignment between the word (tokens) of the sentence and the nodes of the corresponding graph (see, for instance, (Flanigan et al., 2014; Wang et al., 2015; Zhou et al., 2016; Damonte et al., 2017)). However, this anchoring
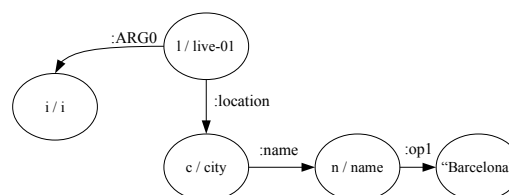

Figure 1: An example of AMR graph for the sentence "I live in Barcelona".

is not provided by manual annotations. Also, the available automatic aligners focus only on the English language (Pourdamghani et al., 2014; Flanigan et al., 2014; Liu et al., 2018), and they do not perform well for other languages. For Portuguese, for instance, the sentence "*Não era surpresa para mim*" (It was no surprise to me), the JAMR aligner (Flanigan et al., 2014) produces alignment only between the token *surpresa* (surprise) and the node `surpresa`, as shown in Figure 2.

```
# ::snt Não era surpresa para mim
# ::alignments 2-3|0

(s / surpresa
    :polarity -
    :domain (e / eu))
```

Figure 2: Alignment produced by JAMR aligner for the sentence "*Não era surpresa para mim*" (It was no surprise to me).

The JAMR aligned only the span `2-3`, which is the token *surpresa* (surprise), with node `0`, which is the root of the graph; `-` and `eu` nodes were not aligned. This wrong or bad alignments occur because of the JAMR aligner adopts a string-match strategy that is focused on the English language. Besides, these issues contribute to a decrease in the performance of AMR parsers. As a result, recent AMR parsing methods have focused on alignment-free approaches (Lyu and Titov, 2018; Zhang et al., 2019a,b). However, they require a large annotated

corpus, which is available only for English.

In this context, aiming to bridge this lack of resources and tools for other languages, we propose an AMR aligner for Portuguese that focuses on a more semantically matched word-concept pair. For that, we use pre-trained word embeddings and the Word Mover's Distance (WMD) function (Kusner et al., 2015) to match span tokens in the sentence with nodes in the graph. Word embeddings capture some semantics information about a corpus, and WMD measures the dissimilarity between two documents even if they have no words in common. With this, it is possible to produce semantically inspired matches instead of only string-match.

To evaluate our approach, we carry out both intrinsic and extrinsic experiments on an annotated corpus from Portuguese. Our aligner produced better alignments than alignment strategies proposed for English and improved AMR parsing for Portuguese, reaching competitive results with an AMR parser designed for that language.

The remaining of this paper is organized as follows. In Section 2, we briefly introduce the related work. Section 3 describes our proposed aligner. In Section 4, we conduct some experiments and evaluations and show our results. Finally, Section 5 concludes the paper indicating future research.

2 Related Work

Flanigan et al. (2014) developed the first AMR aligner, named JAMR. The authors created a rule-based aligner with fourteen heuristic rules to greedily align concepts in the nodes of the graph with tokens in the sentence. The alignment format is a space-separated list of spans with their graph fragments, where each a descriptor specifies each node (e.g. (Gorn, 1965)): 0 for the root node, 0.0 for the first child of the root node, 0.1 for the second child of the root node and so forth. For example, for the sentence, "The boy wants to go.", the JAMR generates alignments according to Figure 3. The JAMR aligned the spans 2-3, 4-5, and 1-2 (that refer to wants, go, and boy, respectively) with the nodes 0, 0.1, and 0.0 (that are the root of the graph, the second child of the root, and the first child of the root, respectively).

Pourdamghani et al. (2014) adopted an unsupervised word alignment technique with machine learning. The authors followed a syntax-based Statistical Machine Translation (SMT) according to the IBM word alignment model (Brown et al.,

```
# ::snt The boy wants to go
# ::alignments 2-3|0 4-5|0.1 1-2|0.0

(w / want-01
    :ARG0 (b / boy)
    :ARG1 (g / go-01
        :ARG0 b))
```

Figure 3: An example of the JAMR aligner for the sentence "The boy wants to go".

1993) to align linearized AMR graphs with English sentences. For the sentence "The boy wants to go", this approach produces alignments, as shown in Figure 4, where '∼ n' specifies a link to the nth English word. As we can see, the third token (wants) was aligned with the concept want-01, the second token (boy) with the concept boy, and the fifth token (go) with the concept go-01.

```
The boy wants to go

(w / want-01 ~ 3
    :ARG0 (b / boy ~ 2)
    :ARG1 (g / go-01 ~ 5
        :ARG0 b))
```

Figure 4: Alignments produced for the sentence "The boy wants to go".

Liu et al. (2018) extended and improved the JAMR aligner by adding semantic resources into the rules, such as GloVe embeddings (Pennington et al., 2014) and the Morphosemantic database (Fellbaum et al., 2009). Besides, they noted that the JAMR aligner requires that words have at least a common longest prefix of four characters, omitting the shorter cases (as word actions aligned with the concept act-01). Thus, their method improved the JAMR aligner in a 4.6% f-score on the LDC2014T2 corpus. The authors also showed that their aligner improved the JAMR (Flanigan et al., 2014) and CAMR (Wang et al., 2015) parsers.

3 Our Aligner

In order to properly adapt AMR parsers from English to Portuguese, we developed an alignment strategy based on document similarity for the Portuguese language. Our method produces alignments in the JAMR aligner format since most of the AMR parsers adopt this alignment type. To support our method, we used the pre-trained GloVe[1] embeddings of 100 dimensions for the Portuguese

---

[1]We also experimented other pre-trained models with dimensions of 50, 100, and 300.

language ([Hartmann et al., 2017](#)) and some lexical resources. We organized our method into three phases over input annotated sentences: preprocessing, mapping, and aligning.

In the first step, we tokenize the sentences and lemmatize each token, applying the Stanza tool ([Qi et al., 2020](#)) trained for Portuguese. The Portuguese tokenization is slightly different from English. For example, some hyphenated words, as "*via-me*" and "*ouvi-la*" (translated for "saw me" and "hear her"), should be separated by the hyphen, whereas other words, as "*segunda-feira*" and "*recém-casados*" (translated for "Monday" and "newly married"), should not be separated. To detail the next steps, we will use Figure 5 as an example.

```
# ::snt Mas Pedro não respondeu
# ::alignments 0-1|0 3-4|0.0 2-3|0.0.0
1-2|0.0.1+0.0.1.0+0.0.1.0.0 5-6|0.0.2 6-7|0.0.2.0

(c / contrast-01                    0
   :ARG2 (r / responder-01          0.0
      :polarity -                   0.0.0
      :ARG0 (p / person             0.0.1
         :name (n / name            0.0.1.0
            :op1 "Pedro"))))        0.0.1.0.0
```

Figure 5: An example of AMR for the sentence "*Mas Pedro não respondeu*"(But Peter did not answer).

In the next step, we mapped each concept to its respective position in the graph. One can see that we mapped the `contrast-01` concept to the root of the graph `0`, its child `responder-01` to `0.0`, and their children `-` and `person` to their respective positions `0.0.0` and `0.0.1`. To do this, we used the Penman tool ([Goodman, 2020](#)).

In the last step, we align the word tokens of the sentence with the concepts of the graph. The AMR language has two concept types: concrete and abstract (or special keywords) ones. The former are those that are explicitly present in the sentence, while the latter are not. In Figure 5, we can see that the `responder-01` is a concrete concept, since it is in the sentence, while the `contrast-01`, `person`, and `name` concepts are abstract[2].

To align concrete concepts, we use the Word Mover's Distance (WMD)[3] ([Kusner et al., 2015](#)) and the pre-trained GloVe embeddings of 100 dimensions. The WMD is a distance function where the lower distance value indicates a higher similarity to the documents. It measures the minimum amount of distance that embedded words of one document need to "travel" to reach the embedded

words of another document.

We used this distance function to evaluate a distance between the embedded word tokens in the sentence and the embedded concepts in the graph to produce alignments with more semantics information than string-match. Furthermore, we empirically defined a maximum distance (threshold) of 1.5 to match a token with a concept, i.e., our strategy maps a word with a concept only if the distance between them is less than the defined threshold. Figure 6 shows this strategy to align the words of the sentence with concrete concepts of the graph.
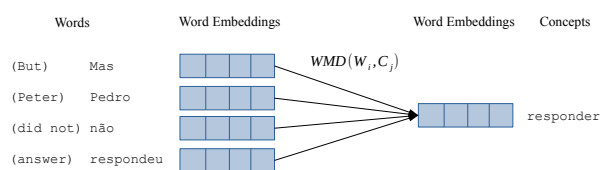

Figure 6: Aligning word tokens with concrete concepts

From this figure, $W = \{w_1, ..., w_n\}$ is the set of words of a sentence and $C = \{c_1, ..., c_n\}$ is the set of concrete concepts of the graph. Our method aligns a w_i with a c_j if and only if the WMD value between w_i and c_j is lower than 1.5, and that value is the lowest among the other delta values.

To align abstract concepts, we use some lexical resources (list of words) to aid and get a higher recall in the alignment. At this time, the AMR formalism has 44 abstract concepts[4] and 110 concepts that represent named entities[5]. For instance, in Figure 5, the `person` is a concept that produces a named entity and, as this concept is in the resource, our alignment strategy aligns this concept and its children with the *Pedro* (Peter) token, which is the span `1-2`, generating the alignment `1-2|0.0.1+0.0.1.0+0.0.1.0.0`, which means that the span `1-2` is aligned with the concept `person` (`0.0.1`) plus `name` (`0.0.1.0`) plus "`Pedro`" (`0.0.1.0.0`).

In addition to named entities and abstract concepts, AMR concepts encompass contrastive conjunctions and negations. To align these concepts, we also created two more lexical resources [6,7] (list of words) for these concepts. Likewise, as ab-

---

[2]"`Pedro`" and `-` are constants, as they get no variable.
[3]We also tested other similarity metrics.

[4]https://amr.isi.edu/doc/amr-dict.html
[5]https://www.isi.edu/~ulf/amr/lib/ne-types.html
[6]https://www.isi.edu/~ulf/amr/lib/popup/contrast.html
[7]https://www.enchantedlearning.com/wordlist/negativewords.shtml

stract concepts and named entities, as the words *Mas* (But) and *não* (not) are in the resources, our alignment method aligns the spans `0-1` and `2-3`, which are the words *Mas* (But) and *não* (not), respectively, with nodes `0` and `0.0.0`, which are `contrast-01` and `-`, respectively (see Figure 5). Our alignment tool is available at `http://github.com/rafaelanchieta/amr-aligner`. In what follows, we detail our experiments with the aligner and the obtained results.

4 Experiments and Results

We performed two experiments, one intrinsic and another extrinsic. In the first, we randomly chose and manually aligned one hundred sentences with their respective AMRs from the Little Prince corpus (Anchiêta and Pardo, 2018a). Then, we compared the manual alignment with the alignments produced by Flanigan et al. (2014) (`JAMR`), Pourdamghani et al. (2014) (henceforth, we refer to it as `UNSU`), Liu et al. (2018) (`TAMR`), and our proposed aligner. We converted the alignment format of the `UNSU` aligner to produce alignments in the `JAMR` aligner format. In Table 1, we show the obtained results in the intrinsic evaluation.

Aligner	Precision	Recall	F-score
JAMR	0.71	0.86	0.78
UNSU	0.48	0.58	0.53
TAMR	0.70	0.88	0.78
OURS	**0.86**	**0.91**	**0.89**

Table 1: Results in the intrinsic evaluation

As we can see, our aligner outperformed those developed for English, which means that our alignment strategy produced alignments more consistent with those manually produced. To get these values, we followed the evaluation method of Flanigan et al. (2014).

In order to confirm the intrinsic evaluation results, we performed an extrinsic evaluation. Thus, we adapted the AMR parsers of Damonte et al. (2017) (henceforth, we refer to it as `AMREager`) and Wang et al. (2015) (henceforth, we refer to it as `CAMR`) for the Portuguese language. We chose these parsers because they use alignments, are open source, need only minor modifications for reuse with other languages, and have a good performance on small corpora.

We trained these parsers on The Little Prince

corpus of the Portuguese language (Anchiêta and Pardo, 2018a), which contains 1,274, 145, and 143 sentences for training, development, and testing, respectively. To compare the results of the parsers, we used the traditional Smatch metric (Cai and Knight, 2013) and the more recently proposed SEMA metric (Anchiêta et al., 2019). SEMA is a more robust metric that considers the parent of the nodes in the graph, producing fairer results than the Smatch metric. Table 2 shows the obtained results in the extrinsic evaluation.

Parser	Aligner	Smatch			SEMA		
		P	R	F1	P	R	F1
CAMR	JAMR	0.46	0.34	0.39	0.20	0.14	0.16
	UNSU	0.35	0.29	0.32	0.15	0.10	0.12
	TAMR	0.47	0.36	0.41	0.24	0.18	0.20
	OURS	**0.54**	**0.47**	**0.50**	**0.32**	**0.27**	**0.29**
AMREager	JAMR	0.44	0.33	0.38	0.18	0.13	0.15
	UNSU	0.34	0.27	0.30	0.14	0.09	0.11
	TAMR	0.38	0.34	0.36	0.17	0.15	0.16
	OURS	**0.51**	**0.45**	**0.48**	**0.30**	**0.26**	**0.28**

Table 2: Results in the extrinsic evaluation

From this table, one realizes that our aligner improved the adapted AMR parsers for Portuguese in both metrics, confirming the results of the intrinsic evaluation. Moreover, the `CAMR` parser achieved a competitive result (50% f-score) compared to the `RBAMR` parser (53% f-score over the same corpus) (Anchiêta and Pardo, 2018b), a rule-based AMR parser designed for the Portuguese language.

We also performed a fine-grained error analysis to identify the weaknesses of our aligner. For that, we used the evaluation tool of Damonte et al. (2017) to compare the `CAMR` parser, as it achieved the best results, with the best aligners `JAMR`, `TAMR`, and `OURS`. We present the results in Table 3.

CAMR	JAMR	TAMR	OURS
Metric	**F-score**		
Unlabeled	0.46	0.48	**0.58**
No WSD	0.41	0.42	**0.52**
NER	0.00	**0.13**	0.00
Wiki	0.00	0.00	0.00
Negations	0.00	0.00	**0.61**
Concepts	0.52	0.53	**0.58**
Reetrancies	0.05	0.06	**0.07**
SRL	0.32	0.37	**0.52**

Table 3: Fine-grained results

We can see that the `CAMR` parser + our aligner outperformed the other aligners in most metrics.

The models tied for the Wiki metric due to the corpus not having wiki annotation. In the NER metric, the TAMR aligner performed better than the other aligners. This result is because of the specific rules used to align named entities and the Morphosemantic database that this aligner makes use of. Besides, our aligner achieved only 0.07 of reentrancies, as the aligner is not prepared to align reentrancies. One solution could be to model reentrancies as a tree, according to Zhang et al. (2019a).

Another issue is the occurrence of hidden subjects in Portuguese, i.e., sentences where the subject 'I' is in the graph, but it is not in the sentence (it is implicit). Our aligner tool also ignores this phenomenon. One solution could be to apply a preprocessing to identify and include the hidden subjects in the sentences.

Treating these issues remain for future work. We also intend to investigate the Morphosemantic database, aiming to improve the accuracy in the alignment of named entities.

The adapted AMR parsers are available at http://github.com/rafaelanchieta/CAMR-PT and http://github.com/rafaelanchieta/amr-eager-pt.

5 Conclusion

In this paper, we presented an AMR alignment method designed for the Portuguese language. It is based on pre-trained word embeddings and Word Mover's Distance to match word tokens in the sentences and nodes in the corresponding AMR graphs. This simple approach may be adopted for other languages with few resources, aiming to get tools for natural language understanding tasks. Furthermore, this aligner may help to build or increase semantic resources, using a promising approach as back-translation (Sobrevilla Cabezudo et al., 2019). Future work includes adopting multilingual word embeddings (Lample et al., 2018) to produce alignments for other languages. More details about AMR resources and tools for the Portuguese language may be found at the OPINANDO project webpage [8].

Acknowledgments

[8] https://sites.google.com/icmc.usp.br/opinando/

References

Rafael Anchiêta and Thiago Pardo. 2018a. Towards AMR-BR: A SemBank for Brazilian Portuguese language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 974–979, Miyazaki, Japan. European Languages Resources Association.

Rafael Torres Anchiêta, Marco A. S. Cabezudo, and Thiago Alexandre Salgueiro Pardo. 2019. SEMA: an extended semantic evaluation metric for amr. In *(To appear) Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France.

Rafael Torres Anchiêta and Thiago Alexandre Salgueiro Pardo. 2018b. A rule-based amr parser for portuguese. In *Ibero-American Conference on Artificial Intelligence*, pages 341–353, Trujillo, Peru. Springer International Publishing.

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.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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.

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. Association for Computational Linguistics.

Christiane Fellbaum, Anne Osherson, and Peter E. Clark. 2009. Putting semantics into wordnet's "morphosemantic" links. In *Human Language Technology. Challenges of the Information Society*, pages 350–358, Berlin, Heidelberg. Springer Berlin Heidelberg.

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. Association for Computational Linguistics.

Michael Wayne Goodman. 2020. Penman: An open-source library and tool for AMR graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 312–319, Online. Association for Computational Linguistics.

Saul Gorn. 1965. Explicit definitions and linguistic dominoes. In *Systems and Computer Science, Proceedings of the Conference held at Univ. of Western Ontario*, pages 77–115.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1989–1993, Las Palmas, Canary Islands - Spain. European Language Resources Association.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 957–966, Lille, France. PMLR.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, Vancouver, Canada.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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*, pages 425–429, Doha, Qatar. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo, Simon Mille, and Thiago Pardo. 2019. Back-translation as strategy to tackle the lack of corpus in natural language generation from semantic representations. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 94–103, Hong Kong, China. Association for Computational Linguistics.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. A transition-based algorithm for AMR parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019a. 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. Association for Computational Linguistics.

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*, pages 3777–3789, Hong Kong, China. Association for Computational Linguistics.

Junsheng Zhou, Feiyu Xu, Hans Uszkoreit, Weiguang Qu, Ran Li, and Yanhui Gu. 2016. AMR parsing with an incremental joint model. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 680–689, Austin, Texas. Association for Computational Linguistics.