From Sentences to Documents: Extending Abstract Meaning Representation for Understanding Documents

De Oraciones a Documentos: extendiendo Abstract Meaning Representation para la comprensión de textos

Paloma Moreda, Armando Suárez, Elena Lloret, Estela Saquete, Isabel Moreno
Department of Software and Computing Systems, University of Alicante
Apdo. de Correos 99 E-03080, Alicante, Spain
{moreda,armando,elloret,stela,imoreno}@dlsi.ua.es

Abstract: The overabundance of information and its heterogeneity requires new ways to access, process and generate knowledge according to the user's needs. To define an appropriate formalism to represent textual information capable to allow machines to perform language understanding and generation will be crucial for achieving these tasks. Abstract Meaning Representation (AMR) is foreseen as a standard knowledge representation that can capture the information encoded in a sentence at various linguistic levels. However, its scope only limits to a single sentence, and it does not benefit from additional semantic information that could help the generation of different types of texts. Therefore, the aim of this paper is to address this limitation by proposing and outlining a method that can extend the information provided by AMR and use it to represent entire documents. Based on our proposal, we will determine a unique, invariant and independent standard text representation, called canonical representation. From it and through a transformational process, we will obtain different text variants that will be appropriate to the users' needs.
Keywords: AMR, documents, canonical representation, user

Resumen: La sobreabundancia de información y su heterogeneidad requieren nuevas formas de acceder, procesar y generar conocimiento de acuerdo con las necesidades del usuario. Por ello, definir un formalismo adecuado para representar la información textual capaz de permitir a los ordenadores comprender y generar el lenguaje, es crucial para lograr esta tarea. Abstract Meaning Representation (AMR) es una representación del conocimiento estándar que puede capturar la información codificada en una oración en varios niveles lingüísticos. Sin embargo, su alcance se limita a una sola oración, y no se beneficia de la información semántica adicional que podría ayudar a la generación de diferentes tipos de textos. En este artículo propondremos un método que amplia la información proporcionada por AMR y la utiliza para representar documentos completos. En base a nuestra propuesta, definiremos una representación de texto estándar única, invariable e independiente, llamada representación canónica. A partir de la cual, y mediante un proceso de transformación, obtendremos diferentes variantes de texto que serán apropiadas para las necesidades de los usuarios.
Palabras clave: AMR, documentos, representación canónica, usuario

1 Introduction

In the context of the Digital Society, the over-abundance of information and its heterogeneity requires new ways to access, process and generate knowledge according to the user needs. In this regard, Human Language Technologies (HLT) play a key role in the analysis, processing and understanding of information. However, the progress made in HLT applications focuses on solving only specific tasks in specific domains, offering solutions from a partial and isolated perspective, without keeping a common model for knowledge extraction, and without considering the user needs as a cross-cutting and intrinsic as-

pect in the process.

Therefore, the main goal of this paper is based on the need for conducting research into a new paradigm for text understanding that will allow us to determine a unique, invariant and independent standard text representation, called canonical representation. From this representation and through a transformational process, we will obtain different text variants that will be appropriate to the users' needs, so that these transformations can be applied to other HLT tasks, such as simplification, enrichment or summarization.

To that end, this paper defines the canonical representation of texts, and how it could be used to generate variations of texts is shown. Such representations are defined using as a basis the Abstract Meaning Representation (AMR) formalism (Banarescu et al., 2013) with improvements: extension of graph at document level and inclusion of additional information and annotation with the VerbNet set of roles (Schuler, 2006).

The remainder of this paper is organized as follows. Section 2 reviews previous work using AMR formalism. Next, Section 3 introduces the canonical representation of texts. Latter, Section 4 shows an example text together with its canonical representation and possible variations. Last, conclusions and future work are outlined.

2 Related Work

Among the specific formalisms for representing natural language at different linguistic levels (lexical, syntactical, semantic, etc.), AMR has gained popularity in the last years since this type of representation can capture semantic aspects of sentences, thus helping Natural Language Understanding and Generation. Although we can find research focused on developing visual tools to better understand AMR annotations (Saphra and Lopez, 2015) or the creation of AMR-annotated corpora (Banarescu et al., 2013; Vanderwende, Menezes, and Quirk, 2015) to be able to train parsers, previous literature has been mostly devoted to automatically address AMR semantic parsing in order to obtain the appropriate representation of a sentence following the AMR guidelines (Vanderwende, Menezes, and Quirk, 2015; Zhou et al., 2016; Goodman, Vlachos, and Naradowsky, 2016; Damonte, Cohen, and Satta, 2017). However, AMR has a great potential for HLT tasks, especially the ones related information generation (e.g., text summarization or natural language generation).

For instance, the use of AMR for text summarization is beneficial for producing abstractive summaries. In this manner, the approach proposed in Liu et al. (2015) partly address this task by building a summary graph from an AMR graph by a concept merging step. In their approach, the coreferent nodes of the graph were merged together. These nodes were either name entities or dates. The authors tested their method in newswire documents and compared the summaries generated from AMR gold-standard annotations with respect to use the output of an AMR parser (in particular JAMR (Flanigan et al., 2014)), and despite being differences in the results obtained, in both cases the results were state-of-the-art according to the summarization task, thus being a very promising method to integrate in abstractive summarization approaches. A similar idea is addressed in Dohare and Karnick (2017), where the authors try to overcome some of the limitations of the approach previously described. Whereas in Liu et al. (2015) a single summary graph from the story graph was extracted, assuming that all the important information from the graph could be extracted from a single subgraph, in Dohare and Karnick (2017), multiple subgraphs are extracted each focusing on information in a different part of the story. In this manner, a few important sentences are first selected and then a summary graph is built from the AMR representation. Finally, an existing text generation from AMR is used to finally produced the resulting summaries. The summarization approach was evaluated and compared to other baselines and approaches, improving the results of Liu et al. (2015) (51.3 vs. 44.3) for the ROUGE-1 F-measure metric.

Apart from text summarization, AMR has also been used to directly generate text. In Flanigan et al. (2016) a statistical method relying on discriminative learning is studied. First, a spanning tree is generated from the AMR and then tree-to-string decoder is applied to generate English, based on the probabilities given by a language model. On the other hand, the approach proposed in Pourdamghani, Knight, and Hermjakob (2016) addressed the problem of AMR-to-text as a

phrased-based machine translation problem. The proposed method learned to linearize AMR tokens into an English-like order. The aim is to induce an ordering function that takes any set of edge labels from AMR as input and produces a permutation of those labels. Several linearization methods were analyzed (e.g., taking into account the most common order for each role in the data, or using different binary classifiers to learn the order for each type of feature). The results achieved overperformed the previous results obtained in Flanigan et al. (2016).

Other formalisms for representing text have been proposed. In Martínez-Barco et al. (2013), a conceptual representation schema was proposed for decomposing natural language into smaller units that could be later combined to generate different types of text (such as summary, an enriched text, or a simplified text), taking into account users' needs. However, it was only a theoretical approach without any implementation, so despite being interesting, it could not be materialized and tested. Taking as a basis this conceptual model, and having analyzed that AMR may be an appropriate specific formalism to represent and generate language, we would like to combine the potentials of both of them by first extending the AMR representation to entire documents and enriching it with additional information, and then being able to generate different types of texts depending on users' needs (e.g. a summary, a simplified text, a schema), thus improving the accessibility of information for any type of user.

3 Canonical representation of texts

Our target is to define a standard representation of a text that allows us to generate different versions of it (summaries, simplifications, enrichments, etc.). In order to do this we are going to use the AMR language. In AMR, each sentence in a text is a single rooted and directed graph, that implies a semantic limited canonical representation of the sentence. Hence, taking AMR as a basis, we propose to enrich each sentence's AMR and use this new information to link as much as possible the sentence level graphs. In this manner, all the semantics of a document is added to express completely the meaning of a text. This would be particularly useful for expressing the meaning of text not only in a different manner, but also to be widely understood by any audience.

Our proposal is performed in two steps. Figure 1 shows the architecture of the system:

1. **Sentence level representation**: Generating AMR per sentence and enriching them with extra information.
   (a) Representing each sentence with AMR formalism using existing parsers, such as JAMR annotator (Flanigan et al., 2014).
   (b) Integrating VerbNet (Schuler, 2006) set of semantic roles, instead of PropBank owing to the fact that the latter is not able to generalize the meaning of the numeric roles, and thus finding cases in which, for instance, a location role for the verb "go" can be represented by ARG2 or ARG4. This problem disappears when the VerbNet set of roles is employed, as AMR originally defined in the PENMAN project (Langkilde and Knight, 1998).
   (c) **Temporal information resolution**. All the temporal expressions in the text will be detected and resolved using TIPSem system (Llorens, Saquete, and Navarro-Colorado, 2013). This system is based on morphosyntactic knowledge plus semantic knowledge, specifically, semantic networks and semantic roles. TIPSem is able to automatically annotate all the temporal information according to TimeML standard annotation scheme, that means annotating all the temporal expressions (TIMEX3), events (EVENT) and links between them.
   (d) Resolving **concepts and entities** in the text is tackled using Babelfy (Moro, Cecconi, and Navigli, 2014). This system addresses entity linking and word sense disambiguation in an unified-manner. Babelfly is a graph-based approach based on a loose identification that selects high-coherence semantic interpretations. It allows the extraction of information related to them, such as synonyms.

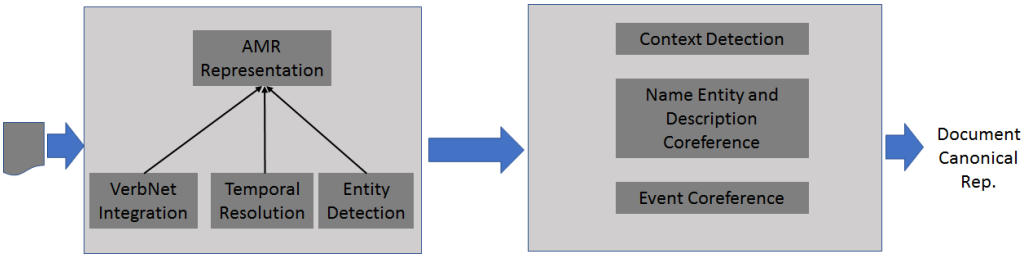

Figure 1: Architecture

2. **Document level representation**: Merging the different extended AMR's of the document by means of: context, name-entity coreference and event coreference.

   (a) Obtaining **context information** such as domain of the text and document creation time. In order to obtain the domain of the text our approach relies on calculating the more frequent domains in a text by agglomerating all the domains linked to all the words with a domain extracted from Babelnet in the previous step, and finally the list of labels is sorted according to the overall frequency.

   (b) **Coreference resolution**. All mentions referring to the same entity are extracted using an state-of-the-art tool, the Standford Coreference Resolution System (Clark and Manning, 2015). It tackles both pronominal and nominal coreference. The latter refers to definite descriptions, which are noun phrases introduced by a definite article and denoting a particular entity. This system implements an statistical mention-ranking model to iterate through each mention in the document to establish a coreference link with a preceding mention.

   (c) **Event Coreference resolution.** All those events mentions referring to the same real fact will be merged in the final graph in one single node in order to relate the different AMR's and simplify the graph of the whole document. Event coreference will be determined in a two clustering process. First, a temporal clustering will be performed, so all the events happening at the same time will be clustered together. After this, a semantic clustering is performed so the events are clustered using lexical semantics (lemmas and synonyms) and distributional semantic knowledge (word2vec) in order to resolve event coreference (Navarro-Colorado and Saquete, 2016). The temporal clustering is not a trivial task, but we are using the temporal information annotated by TipSEM system for this purpose. According to Tempeval-3 (UzZaman et al., 2013) evaluation, TipSEM system is obtaining an F1-score of 65.31% at temporal expression performance and an F1-score of 42.39% at temporal awareness regarding temporal relations. Regarding event coreference, we are using the system described at Navarro-Colorado and Saquete (2016), that combines both tasks with a final F1-score of 26.5% for the experiments involving temporal, lexical and distributional clustering, which improves the current state-of-the-art systems and shows a significant advance in the Cross-Document Event Ordering task.

3.1 The canonical formalism

In order to build a complete text graph from the AMR representation of each individual sentence, and considering the fact that all the nodes in AMR representation has a variable, once the coreference is resolved, all the nodes referring to the same thing (entities or events) will be identify with the same variable. Apart from this, our extended AMR representation will introduced these *new* relations or edges to the nodes (when necessary):

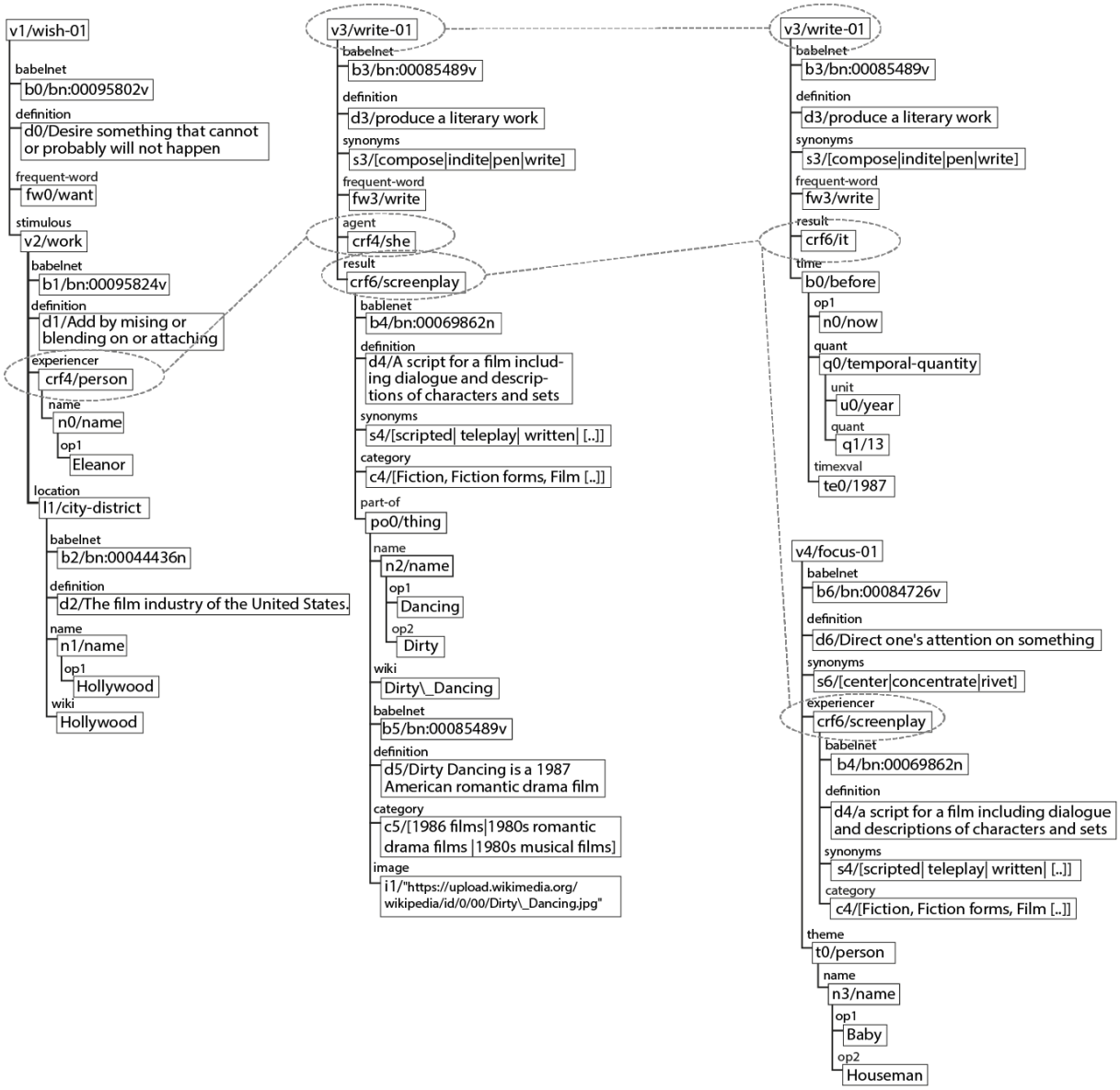

Figure 2: Four sentences AMR graphs with merging points

(a) *:dct* – document creation time;

(b) *:topic* – document topic;

(c) *:timexval* – ISO temporal value of temporal expressions;

(d) *:category_related* – category of the concept;

(e) *:domain* – ID of BabelDomain;

(f) *:babelnet* – ID from BabelNet for each concept or entity;

(g) *:definition* – explanation for a term;

(h) *:category* – the semantic class associated to the term;

(i) *:image* – the image associated to a concept or entity;

(j) *:synonyms* – alternative terms for the same concept or entity;

(k) *:frequent-word* – synonym that is most commonly used.

4 Case Example

This section shows a possible output from a text example formed by four sentences, as well as possible transformations derived from the original text.

Typically, an AMR representation in PENMAN notation would give us something similar to the next output for the sentence "*Eleanor wished to work in Hollywood.*", including some of the enrichments mentioned before:

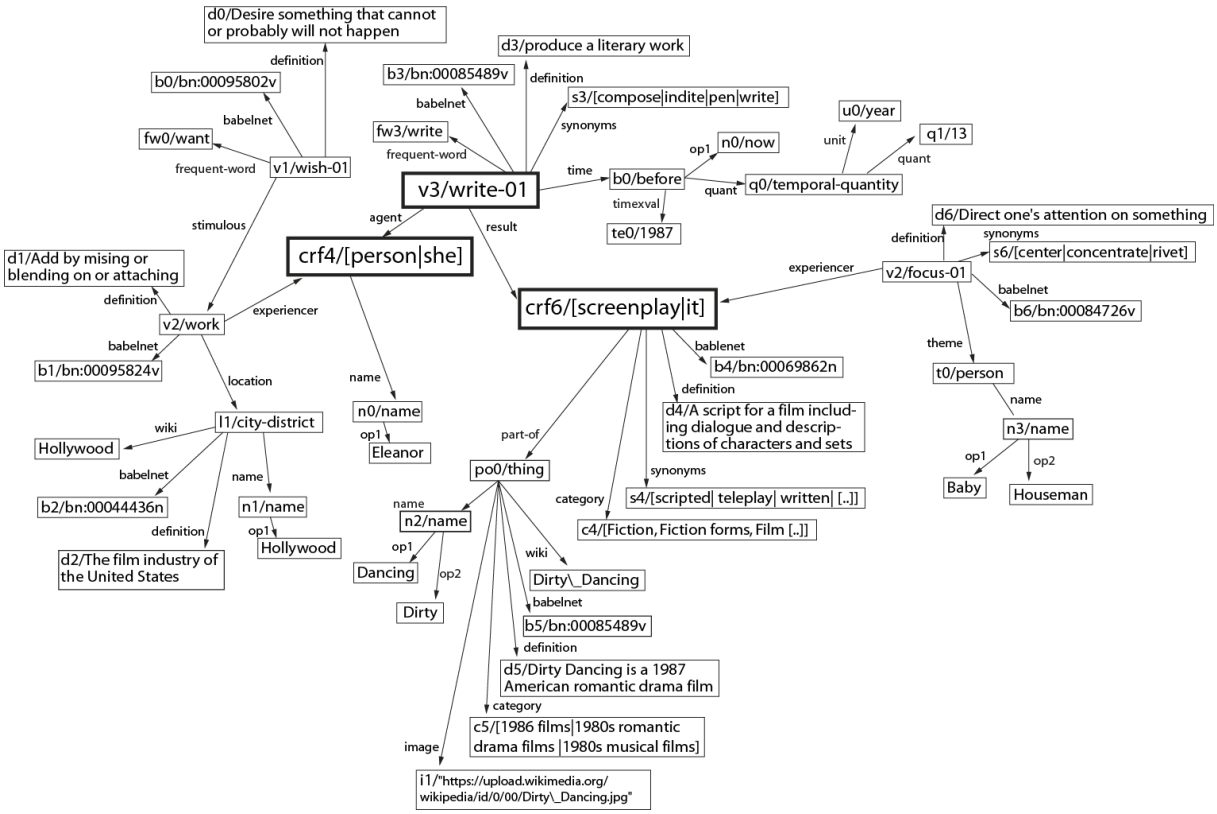

Figure 3: Whole text graph after merging

```
%Eleanor wished to work in Hollywood.
(v1 / wish-01
    :babelnet (b0 / bn:00095802v)
    :definition (d0 / "Desire something that
    cannot or probably will not happen")
    :frequent-word (fw0 / "want")
    :stimulous (v2 / work-01
         :babelnet (b1 / bn:00095824v)
         :definition (d1 / "Add by mising or
            blending on or attaching")
         :experiencer (crf4 / person
            :name (n0 / name :op1 "Eleanor"))
         :location (l5 / city-district
            :babelnet(b2 / bn:00044436n)
            :definition (d2 / "The film
               industry of the United States.")
            :name (n1 / name :op1 "Hollywood")
            :wiki "Hollywood")))
```

Suppose a text with these four sentences: "*Eleanor wished to work in Hollywood. She wrote the screenplay of Dirty Dancing. It was written thirty years ago. The screenplay was focused on Baby Houseman.*". Given such input four direct graphs can be produced such as those shown in Figure 2. We also mark the nodes of the graphs that represent the same concept. These are merging points which allows us to produce the final canonical representation of the whole text.

For example, (*v3/write*) is an AMR node with a variable v3 appearing in two of the sentences, as well as the pairs (*crf4/person – crf4/she*) and (*crf6/it – crf6/screenplay*), labeled by co-reference resolution. Using these merging points a final graph is shown in Figure 3.

Using the canonical representation, the text could be transformed without losing its meaning through the navigation of the AMR text graph. Each transformation could be appropriate in a particular situation or for a specific purpose. Figure 4 shows two possible transformations for our example. The first transformation is an extended text containing the explanation for the term "screenplay" and using synonyms of the original words, whereas the second shows a headline. The former could be useful to someone who is not an expert or has reading comprehension difficulties.

5 Conclusions

In this paper, a text representation using Abstract Meaning Representation (AMR) is proposed. Originally, AMR is intended to represent the concepts and their relationships of one sentence only. In this manner, the set of sentences that compose an entire text results in a set of disconnected AMR graphs. Our proposal consists of an architecture and

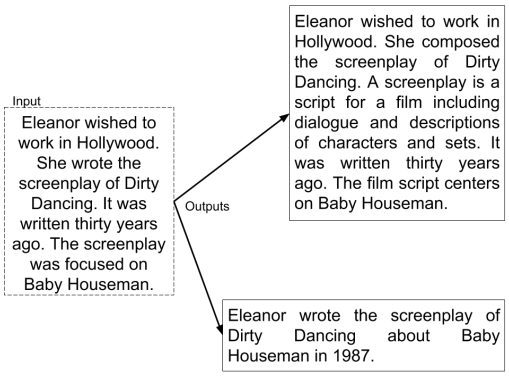

Figure 4: Examples of possible inflexions

a method to add new data to offer more semantic information at sentence level, but to link and merge such graphs too. The final goal is to achieve a unique, invariant and independent standard representation of entire documents. Such canonical representation will allow us to generate new variants of the analyzed text such us summaries, simplifications, etc. to satisfy different user's needs.

In this paper we illustrate this enrichment over four sentences representing a short document, showing the resulting AMR graphs and how they are merged and, finally, an example of text variants that could be generated from this text canonical representation.

In order to do this, several NLP tools were used to enrich the basic AMR representation such as semantic role labelling, entity identification and resolution, temporal information resolution, document categorization and coreference resolution. These added variables and values permit to find linking points on graphs in order to relate as much sentences as possible, as the intuitive notion of what a document is seem to point out. Since the errors made by these NLP tools may affect the accuracy of the information to be represented in the AMR graph, a validation process will be done to detect wrong information with the purpose of avoiding creating a noisy AMR graph. This could be done by first using available annotated corpora as input for the AMR graph and then, comparing the result with the one obtained when using the automatic tools.

In the immediate future, we will focus on analyzing new sources of information to add and managing user's profiles to offer the most useful text variants for them. For the evaluation of our final system, internal metrics will be provided, and a pedagogical point of view will be explored in the form of automatic assessments generation.

Acknowledgments

Research partially supported by the Spanish Government (grants TIN2015-65100-R; TIN2015-65136-C02-2-R).

References

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Clark, K. and C. D. Manning. 2015. Entity-centric coreference resolution with model stacking. 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 1405–1415, Beijing, China, July. Association for Computational Linguistics.

Damonte, M., S. B. Cohen, and G. 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, April. Association for Computational Linguistics.

Dohare, S. and H. Karnick. 2017. Text summarization using abstract meaning representation. *CoRR*, abs/1706.01678.

Flanigan, J., C. Dyer, N. A. Smith, and J. Carbonell. 2016. Generation from abstract meaning representation using tree transducers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–739, San Diego, California, June. Association for Computational Linguistics.

Flanigan, J., S. Thomson, J. Carbonell, C. Dyer, and N. A. Smith. 2014. A dis-

criminative 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, June. Association for Computational Linguistics.

Goodman, J., A. Vlachos, and J. Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Berlin, Germany, August. Association for Computational Linguistics.

Langkilde, I. and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *COLING-ACL '98, Proceedings of the Conference*, pages 704 – 710, Montreal, Canada, August. Association for Computational Linguistics.

Liu, F., J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado, May–June. Association for Computational Linguistics.

Llorens, H., E. Saquete, and B. Navarro-Colorado. 2013. Applying Semantic Knowledge to the Automatic Processing of Temporal Expressions and Events in Natural Language. *Information Processing & Management*, 49(1):179–197.

Martínez-Barco, P., A. F. Rodríguez, D. Tomás, E. Lloret, E. Saquete, F. Llopis, J. Peral, M. Palomar, J. M. G. Soriano, and M. T. Romá-Ferri. 2013. LEGOLANG: técnicas de deconstrucción aplicadas a las tecnologías del lenguaje humano. *Procesamiento del Lenguaje Natural*, 51:219–222.

Moro, A., F. Cecconi, and R. Navigli. 2014. Multilingual word sense disambiguation and entity linking for everybody. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, ISWC-PD'14, pages 25–28, Aachen, Germany, Germany. CEUR-WS.org.

Navarro-Colorado, B. and E. Saquete. 2016. Cross-document event ordering through temporal, lexical and distributional knowledge. *Knowl.-Based Syst.*, 110:244–254.

Pourdamghani, N., K. Knight, and U. Hermjakob. 2016. Generating english from abstract meaning representations. In *Proceedings of the 9th International Natural Language Generation conference*, pages 21–25, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

Saphra, N. and A. Lopez. 2015. Amrica: an amr inspector for cross-language alignments. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 36–40, Denver, Colorado, June. Association for Computational Linguistics.

Schuler, K. K. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

UzZaman, N., H. Llorens, L. Derczynski, J. Allen, M. Verhagen, and J. Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. ACL. ISBN: 978-1-937284-49-7.

Vanderwende, L., A. Menezes, and C. Quirk. 2015. An amr parser for english, french, german, spanish and japanese and a new amr-annotated corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, Denver, Colorado, June. Association for Computational Linguistics.

Zhou, J., F. Xu, H. Uszkoreit, W. QU, R. Li, and Y. 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, November. Association for Computational Linguistics.