Towards Universal Semantic Representation

Huaiyu Zhu
IBM Research - Almaden
650 Harry Road,
San Jose, CA 95120
huaiyu@us.ibm.com

Yunyao Li
IBM Research - Almaden
650 Harry Road,
San Jose, CA 95120
yunyaoli@us.ibm.com

Laura Chiticariu
IBM Watson
650 Harry Road,
San Jose, CA 95120
chiti@us.ibm.com

Abstract

Natural language understanding at the semantic level and independent of language variations is of great practical value. Existing approaches such as semantic role labeling (SRL) and abstract meaning representation (AMR) still have features related to the peculiarities of the particular language. In this work we describe various challenges and possible solutions in designing a semantic representation that is universal across a variety of languages.

1 Introduction

Natural languages have many syntactic variations for expressing the same meaning, not only within each language but more so across languages, making syntactical analysis cumbersome to use by downstream applications. Semantic understanding of natural language is fundamental for many applications that take natural language texts as part of their input. Semantic Role Labeling (SRL) analyzes the predicate-role structure at the shallow semantic parsing level (e.g., PropBank (Kingsbury and Palmer, 2002)). At a deeper level, Abstract Meaning Representation (AMR) provides a rooted, directional and labeled graph representing the meaning of a sentence (Banarescu et al., 2013), focusing on semantic relations between concepts such as PropBank predicate-argument structures while abstracting away from syntactic variation.

Many applications require multilingual capabilities, but SRL and AMR annotation schemes designed for individual languages have language-dependent features. For example, Hajic et al. (2014); Xue et al. (2014) observed AMRs designed for different languages have differences, some accidental but others are more fundamental. Several efforts are underway to create more cross-lingual natural language resources. Universal Dependencies (UD) is a framework for cross-linguistically consistent grammatical annotation.(De Marneffe et al., 2014). The Universal Proposition Banks project aims to annotate text in different languages with a layer of "universal" semantic role labeling annotation, by using the frame and role labels of the English Proposition Bank to label shallow semantics in sentences in new target languages(Akbik et al., 2015). Similarly, Damonte and Cohen (2018) use AMR annotations for English as a semantic representation for sentences written in other languages, utilizing an AMR parser for English and parallel corpora to learn AMR parsers for additional languages.

Despite these efforts, some remaining inter-language variations important for practical usage are not yet captured by the efforts so far. They create obstacles to a truly cross-lingual meaning representation which would enable the downstream applications be written for one language and applicable for other languages. The purpose of this paper is two-fold. One objective is to highlight some of these remaining issues and call the attention of the community to resolving them. Another objective is to advocate a form of abstract meaning representation geared towards cross-lingual universal applicability, in the same spirit of AMR but somewhat simplified, with the following major similarities and differences

- Like AMR, it makes use of PropBank style predicate-argument structures.
- It does not have AMR style concept nodes. It does not infer relations among instances and concepts other than those expressed explicitly, nor perform co-reference resolution.
- It is geared towards cross-lingual representation of logical structures, such as conjunctions and conditionals.
- It assigns features to nodes, to promote structural simplicity and to increase extensibility.

Proceedings of the First International Workshop on Designing Meaning Representations, pages 177–181
Florence, Italy, August 1st, 2019

We will illustrate, through several examples, the kinds of issues that arise from attempting to create a universal meaning representation, and the challenges in resolving these issues. We will describe our tentative solutions and call the attention of the community to these issues.

2 Examples of semantic variations

Across different languages, semantic structures are much more uniform than syntactic structures. However, there are still language variations in shallow semantics. In this section we look at a number of examples.

2.1 Temporal semantics

Predicates often represent actions that happen, or states or properties that exist or change, in a certain time frame. Different languages have different ways to express such temporal relations. In English, auxiliary verbs and main verbs are usually combined with morphological change to express tense and aspect. For example,

*I **am going to speak** to him.* (future-simple)
*I **have spoken** to him.* (present-perfect)
*I **was speaking** to him.* (past-progressive)

Similar meanings are represented differently in other languages. For example, German usually does not distinguish verbs between present-perfect and the past-simple as English, even though it formally has corresponding syntactic structures. Instead the distinction is implied by temporal arguments such as a prepositional phrases. In Chinese the corresponding concepts are represented by adverbs and participles. For example,

*I **have been reading** for a week.* (pres.-perf.-prog.)
*Ich **lese** seit einer Woche.* (past-simple)
我**已经读了**一个星期了. (adverb-verb-participle)

A more abstract representation for tense should be able to unify all these variations. Among available linguistic theories, Reichenbach (1947)'s theory of tense covers a large proportion of these variations. It consists of three points in time: point of event (E), point of reference (R) and point of speech (S), and two ordering relations: anteriority and simultaneity among these points. In English, the relation between S and R corresponds to tense, and the relation between R and E corresponds to aspect. For example, "E–S,R" corresponds to present-perfect and "E,R–S" corresponds to past-simple. The "progressive" aspect is not represented in this framework. It can be added as an additional property. In related work, Smith (1997) provides a richer semantics by regarding temporal aspects as relations among time intervals. TimeML (Saurı et al., 2006) defines a rich variety of time related concepts.

2.2 Expressing modality

In English, modal verbs are auxiliary verbs that express various likelihood of the main verb. These include certainty and necessity (must, shall), intention (would), ability or possibility (can, may, might), etc. Additional idiomatic expressions provide similar functionality. For example,

is capable of, used to, had better to, is willing to

AMR represents syntactic modals with concepts like possible-01, likely-01, obligate-01, permit-01, recommend-01, prefer-01, etc. This English-inspired classification of modality must be extended for other languages. For example, in Chinese the modal verbs include at least the following: 能 (can, may), 会 (can, will, be able to), 要 (want, wish, intend to), 肯 (be willing to, consent), 敢 (dare), 可能 (may), 可以 (can, be allowed to), 应该 (should), 愿意 (be willing to). When combined with negation, these also include 不愿意 (be reluctant to, be unwilling to), etc. There is no compelling reason, other than English convention, that modality has special relation to modal verbs. Considerations of additional languages will likely further extend types of such meanings as well as further refine these meanings.

A cross-lingual framework must allow for all these variation, while providing basic features that allow easy categorization of them. In analogy of Reichenbach's theory of tense, we propose to categorize the modality by considering several dimensions that jointly affect the likelihood of an action:

- Probability or certainty
- Requirement or obligation
- Advisability, recommendation or suggestion
- Ability, capability or permit
- Desire or hope
- Willingness or intention

Each modality expression may have values in one or more of these dimensions.

2.3 Conditionality

The most basic language construct expressing "if A, then B" probably exists in most languages with syntactic variations. For example, in English it is more natural to say "if A, B" or "B if A". Syntactical differences aside, such structures essentially express a relation of two things, A as antecedent and B as consequent. Natural languages can also express, but often not in the same sentence, the more complete structure "if A, then B, else C", There does not appear to be a generally adopted linguistic term for the C part.

Unlike formal logic, natural language often associates additional mood, modality and temporal element with these expressions

X only if Y
X as long as Y
If it were not you, it would not have ....
Had I known it, I would have ...

In English, the subjunctive mood is often associated with conditional structures in making counterfactual assumptions. The term subjunctive corresponds to several different concepts in different languages. For example, in Spanish, the subjunctive can be used with verbs for wishes, emotions, impersonal expressions, recommendations, doubt, denial, hope and other verbs to express what is essentially modality. To accommodate such variations across different languages, one possible design is to consider the two aspects of conditionalality expressions separately. One aspect deals with the logical implication $A \rightarrow B$. The other aspect is to assign tense and modality to the conditionals. The tense can be useful for expressions like "Do A until B", and the modality assigned to the conditional can be used to express the modality associated with the conditional itself, not to the antecedent or consequent.

3 A framework for cross-lingual meaning representation

The refined meanings discussed in previous section must be expressed in a certain framework. SRL does not have sufficient abstract structures for this task. AMR is a better candidate, but we have found it lacking in two aspects. On the one hand, it has a substantial amount of extra information that is neither explicitly expressed in the sentence nor required by downstream applications. On the other hand, it still lacks sufficient structure to express the refined meanings discussed above.

We propose a meaning representation that attempts to simplify AMR while allowing easy incorporation of additional features. The proposed representation is a graph with a small number of node types, flexible features on the nodes, and labeled and directed connections among the nodes. It is not necessarily a tree.

3.1 Nodes

We consider the following types of nodes:

Predicate A predicate in the sense of PropBank

Role A core argument, such as A0, A1, etc., in the sense of PropBank.

Context A non-core argument, such as AM-TMP, AM-LOC, etc. in the sense of PropBank.

Conditional Representation of "if-then-else" structure, including variations like "unless", "as long as", "whenever".

Conjunction Representation of "and", "but", "or", etc. Linguistic conjunctions include "and", "but", "or", "nor", etc. Like AMR, it includes both conjunctions and disjunctions as well negated expressions in terms of logic.

Relational Representation of a linguistic relation among entities that is usually expressed in English with prepositions such as "in", "on", "under", or similar structures representing possessive (e.g, "A's B" vs "B of A").

3.2 Features

Each node is associated with additional features specific to the node type. For example, a Predicate node is associated with features such as the verb sense (eg."speak.01"), as well as tense, modality, polarity, etc.

3.3 Edges

The nodes are connected by edges with well defined types

- Role and Context nodes are connected to Predicate nodes with SRL labels. Context might also be connected to other nodes, such as Conditional, as discussed above.
- A Conditional node is connected to an antecedent node and a consequent node, and optionally to an "else" node.
- A Conjunction node is connected to its constituents.

- A Relation node is connected to its constituents.

3.4 Example representation

An example can illustrate various aspects of this framework. Consider the sentence

> *Had I studied harder last year, I would have been able to pass the exam by the end of the winter and got an A.*

This sentence is constructed so that it can be used to illustrate the issues discussed in this paper.

We will express the graph by describing the nodes and their features. We use Json style notation for features as key-value pairs. Some of the values are literal values, others are references to other nodes, essentially representing the edges with labels. In this example, for the sake of exposition, we will use features that correspond more closely to conventional English linguistic features. For example, Predicates have features tense, aspect, modality and polarity.

A = Conditional {mood: conterfactural, antecedent: B, consequent C }.
B = Predicate {sense: study.01, tense: past, aspect: simple, polarity: positive, modality: normal}.
B_1 = Role {content: I, predicate: B, type: A0 }.
B_2 = Context {content: harder, predicate: B, type: AM-MNR }.
B_3 = Context {content: last year, predicate: B, type: AM-TMP }
C = Conjunction {type: and, members: [C_1, C_2] }
C_1 = Predicate {sense: pass.07, tense: past, aspect: perfect, polarity: positive, modality: ability}.
C_2 = Predicate {sense: get.01, tense: past, aspect: perfect, polarity: positive, modality: ability}.
C_{11} = Role {content: I, predicate: C_1, type: A0 }.
C_{12} = Role {content: exam, predicate: C_1, type: A1 }.
C_{13} = Context {content: by the end of the winter, predicate: C_1, type: AM-TMP }
...

Note the following points:

- The structure of this graph is simpler than AMR graph, mostly by virtue of removing the AMR concept nodes.
- For the remaining nodes the edges connecting them are similar to those in AMR graphs.
- The nodes are typed. Each type has a specific set of features.

Although we have used more traditional feature sets in this example, it is obvious that more orthogonal feature designs as discussed in the previous section can be used instead, without changing the overall structure of the graph.

3.5 Learning features from data

Using techniques similar to those used to transfer SRL and AMR from one language to another(Akbik et al., 2015; Damonte and Cohen, 2018), it is possible to transfer labeling schemes for the additional fewatures and structures discussed in this paper from one language to another. The cross-lingual transfer may also help to discover better feature sets from data. For example, by analyzing equivalent sentences in different languages, it is possible to discover additional candidates for modalilty or better classification of modality. Akbik et al. (2016) showed that it is possible to use correspondences between verb senses in two languages to discover the duplication and aliasing of verb senses. Similar techniques can be applied to verb features such as tense and modality, as well as structural featues such as conditional and relational features. It is our hope that this framework provides a sufficiently versatile scafolding for the community to work together towards a more complete cross-lingual representation of meanings.

4 Conclusions

Creating a universal semantic representation that works across a large number of languages is an important objective for the NLP community. In this paper we described our attempts towards this goal, highlighting the issues and challenges that arise from such efforts. In particular, we described specific issues related to representing tense and modality of predicates, as well issues for expressing relational structures among the entities and predicates. We also present a framework for creating an overall structure to hold the cross-lingual semantics. It is similar to AMR but with a different emphasis. Instead of identifying all the intricate relations among the constituents of a sentence as well as the concepts they correspond to, this representation is aimed at expressing the essential structures and important features of these structures in a cross-lingual fashion. As such it sacrifices certain capabilities of AMR (such as concepts and variables) while emphasizing others (such as defining the features for various node types). It is our hope that this framework can stimulate the community to make progress on the design issues for various features of these structures, and we call upon the community to work together to refine this framework.

References

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

Alan Akbik, Xinyu Guan, and Yunyao Li. 2016. Multilingual aliasing for auto-generating proposition banks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3466–3474.

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.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual abstract meaning representation parsing. In *Proc. 2018 NAACL-HLT*, pages 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–92.

Jan Hajic, Ondrej Bojar, and Zdenka Uresova. 2014. Comparing Czech and English AMRs. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 55–64.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *LREC*, pages 1989–1993.

Hans Reichenbach. 1947. *Elements of Symbolic Logic*. Macmillan & Co, New York.

Roser Saurı, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML annotation guidelines version 1.2. 1.

C.S. Smith. 1997. The parameters of aspect.

Nianwen Xue, Ondrej Bojar, Jan Hajic, Martha Palmer, Zdenka Uresova, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of english amrs to chinese and czech. In *LREC*, volume 14, pages 1765–1772. Reykjavik, Iceland.