Semantic Annotation Issues in Parallel Meaning Banking

Johan Bos
University of Groningen
johan.bos@rug.nl

Abstract

If we try to align meaning representations of translated sentences, we are faced with the following problem: even though concepts and relations ought to be independent from specific natural languages, the non-logical symbols present meaning representations in usually resemble language-specific words. In faithful translations, such symbols can be easily aligned. In informative translations (where more information is provided by the target translation), symbols can be aligned by a symbol denoting an inclusion relation. In loose translations, we need a third combinator to combine symbols with similar but not identical meanings. We show how this can be done with several concrete, non-trival English-German translation pairs. The resulting formalism is a first step towards constructing parallel meaning banks.

Keywords: Semantic Representations, Meaning Banking, Parallel Corpora

1. Introduction

The ingredients of meaning representations can roughly be divided into two categories: the logical symbols, and the non-logical symbols. To the first category belong the quantifiers, the variables, and the boolean operators (negation, conjunction). The members of the second category, the non-logical symbols, are based on the language that is undergoing semantic analysis. For example, a meaning representation for a simple sentence like "John doesn't smoke" would contain the logical symbols $\neg$ and $\wedge$, several variables, and the non-logical symbols JOHN (representing the entity referring to John) and SMOKE (representing the event of smoking). But now suppose I have a good translation of this English sentence into, say, German or Dutch. Arguably, the meaning representation for this translation should not differ a great deal. But what would it look like precisely?

One possible solution is to take a (neutral) auxiliary language for defining the vocabulary of non-logical symbols. But soon one will discover that this option isn't feasible. In natural (non-literal) translations, the source is sometimes more general, sometimes more specific than the target translation. This information will be lost when one relies on a single language. Moreover, phrasal translations will be hard to capture by a single language of symbols.[1] The alternative, and one that will be explored in this paper, is to combine the non-logical symbols of the source and target of a translated sentence into a single meaning representation.

In order to investigate this possibility, we follow a strongly data-driven method. We take non-trivial translation examples from an existing corpus (see Figure 1) and produce the meaning representations for each language. Then we will compare the respective meaning representations, and examine how we could align the two representations. Here we will just consider pairs of English-German translations — the choice for these two close languages makes sense for a pilot study of this kind.

We employ Discourse Representation Theory, DRT (Kamp and Reyle, 1993), as the formal theory of meaning, mainly because it is well-known among semanticists and has covers many linguistic phenomena, but we would like to emphasize that any meaning representation with variables and n-place relations could have been adopted to integrate the ideas put forward in this paper. We will introduce new machinery for representing parallel meanings. We will bring three new operators into play for combining non-logical symbols dealing with faithful translations, informative translations, and loose translations. To make this more readable, we just assume that the non-logical symbols represent the right sense of the concepts expressed by the surface strings. We also assume that each non-logical symbol carries the information of its source language (here: English or German), but don't explicitly show it in the meaning representations for reasons of clarity.

2. Faithful Translations: $\equiv$

Faithful translations are among the easiest to align, because they are often based on word-by-word translations. Consider the examples and corresponding meaning representations given below in Example 1. Here, and in the examples that follow, we show the meaning representation for an English expression and one its German translation, and a parallel meaning representation comprising both source and target language. The mono-lingual meaning representations also show the mappings of discourse referents to surface strings (where dotted variables indicate substitutions that need to take place) for the reader's convenience.

EXAMPLE 1

x p
x ↦ "the chance to ṗ"

CHANCE(x)
TO(x,p)

x p
x ↦ "die Gelegenheit zu ṗ"

GELEGENHEIT(x)
ZU(x,p)

↘ ↙

x p
CHANCE≡GELEGENHEIT(x) TO≡ZU(x,p)

[1] Although there are initiatives, notably the Abstract Meaning Representation project (Banarescu et al., 2013), pursuing closely related goals.

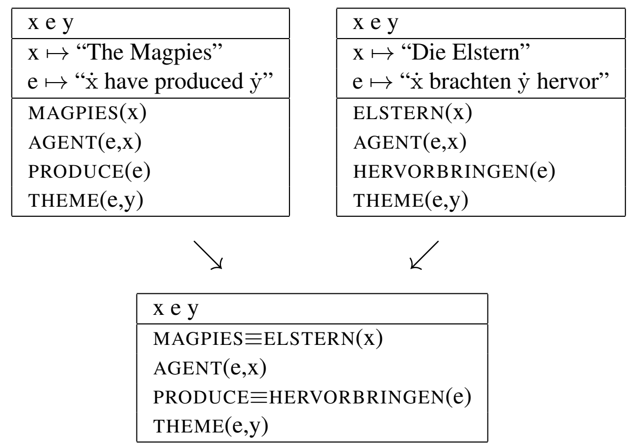
English (en)	**German** (de)
Pubs also provide good value for money, the chance to taste a pint of beer and have a chat with the locals.	Pubs bieten auch ein gutes Preis-Leistungsverhältnis, die Gelegenheit ein Glas Bier zu trinken und mit den Einheimischen zu plaudern.
The “Magpies”, Newcastle United Football Club, have produced some of Britain’s finest players.	Die “Elstern”, wie der Newcastle United Football Club auch genannt wird, brachten einige der besten Fußballspieler Großbritanniens hervor.
Due to the possibility of animals and birds bringing disease to the UK, bringing them with you on holiday is not recommended.	Da Haustiere und Vögel Krankheiten nach Großbritannien einschleppen können, wird davon abgeraten, sie mit in die Ferien zu nehmen.

Figure 1: Examples considered in this study. Source: The English-German Translation Corpus, `http://ell.phil.tu-chemnitz.de/`.

This example illustrates a faithful, literal translation, and as a pleasant consequence there is a simple one-to-one mapping between the non-logical symbols of the source and target language. To arrive at a parallel meaning representation, we combine the non-logical symbols (with the same arity) originating from different languages by simply concatenating them with the help of a new operator: $\equiv$. For instance, the German-originating two-place relation ZU and the English-originating two-place relation TO are combined to yield a new compound non-logical symbol TO$\equiv$ZU.

Now consider Example 2, illustrating some basic neo-Davidsonion event structure.[2] It makes sense to assume that the thematic roles are universal and therefore language independent. Therefore it is not necessary to align the conditions for the roles in the parallel meaning representation: they are shared. However, it could be the case that there are languages that explicitly express a role (for instance, by a preposition), in which case the non-logical symbol denoting that role could be based on it.[3]

EXAMPLE 2



We will give meaning to this new operator by extending a translation function from the meaning representation to first-order logic, $[.]^{fol}$, on the same lines as earlier work in Discourse Representation Theory (Bos, 2004; Kamp and Reyle, 1993). We can define $\equiv$ as follows:

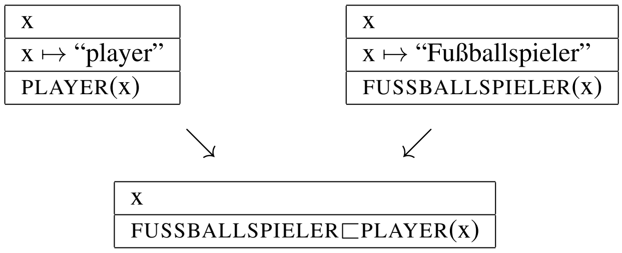
$$[S_i{\equiv}S_j(\mathrm{x}_1,\ldots,\mathrm{x}_n)]^{fol} = S_i(\mathrm{x}_1,\ldots,\mathrm{x}_n) \wedge \forall \mathrm{u}_1,\ldots,\mathrm{u}_n(S_i(\mathrm{u}_1,\ldots,\mathrm{u}_n) \leftrightarrow S_j(\mathrm{u}_1,\ldots,\mathrm{u}_n))$$

This simply says that all these symbols are synonyms, and applied to n of variables, result in logically equivalent meanings. One could compare this to a WordNet (Fellbaum, 1998) synset: given the compound symbol A$\equiv$B, then A and B belong to the same cross-lingual synset.

3. Informative Translations: $\sqsubset$

Translations, however, are rarely as literal and faithful as the previous examples suggest. Consider for instance Example 3, where the English noun “players” is translated into German with the more specific “Fußballspieler”. Even though it is clear from the context in the English sentence that we talk about players that practice the game of football, it isn’t stated explicitly. It would therefore be wrong to align the meanings of these words with the $\equiv$ operator. What we propose to do instead is introducing a new operator, $\sqsubset$, that combines two symbols and specifies that the first is more specific (carries more information) than the second.

EXAMPLE 3



As can be seen in the parallel meaning representation in Example 3, we specified that FUSSBALLSPIELER is more informative than PLAYER. This seems to be a common phenomenon in translation. What’s left to do is giving a formal definition for $\sqsubset$, and we define it in first-order logic as:

$$[S_i{\sqsubset}S_j(\mathrm{x}_1,\ldots,\mathrm{x}_n)]^{fol} = S_i(\mathrm{x}_1,\ldots,\mathrm{x}_n) \wedge \forall \mathrm{u}_1,\ldots,\mathrm{u}_n(S_i(\mathrm{u}_1,\ldots,\mathrm{u}_n) \rightarrow S_j(\mathrm{u}_1,\ldots,\mathrm{u}_n))$$

[2] For simplicity we assume that proper names introduce one-place relations.

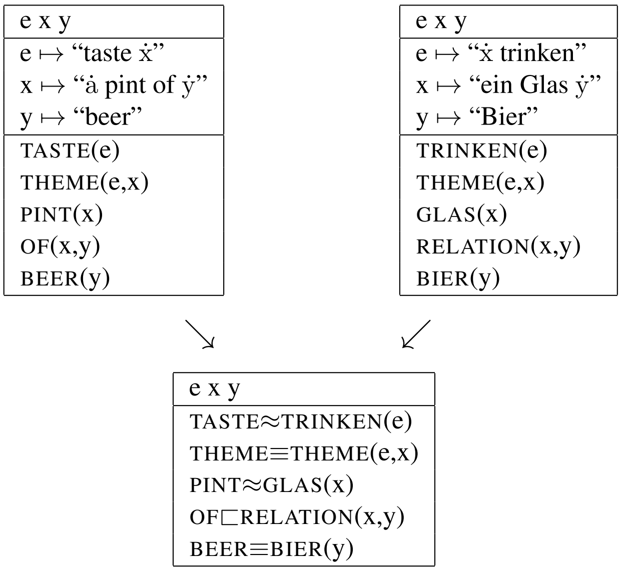
[3] An example that comes to mind is the passive construction in English, where the *agent* role is marked by the preposition “by”. A further example is the semantic role of *recipient* expressed by the preposition *to*, in constructions like “Mary gives the book to John”. See also Example 6.

For instance, given the compound symbol A⊏B applied to x, then A(x) holds, and if A(x) holds then also B(x) holds. In the parlance of WordNet (Fellbaum, 1998) practitioners, A would be a hyponym of B.

4. Loose Translations: ≈

An old proverb says that a translation cannot be both faithful and beautiful. Loose translations often just sound better. A case in point is "taste a pint of beer" and its German rendering "ein Glas Bier trinken": a pint (a unit of measurement) isn't the same as a glass (a container), and tasting isn't the same as drinking, although in WordNet (Fellbaum, 1998) they are both co-troponyms of *consume*. To align such loose translations we propose a new operator for symbol alignment: ≈, illustrated by Example 4.

EXAMPLE 4



```
e x y                      e x y
e ↦ "taste ẋ"              e ↦ "ẋ trinken"
x ↦ "ȧ pint of ẏ"          x ↦ "ein Glas ẏ"
y ↦ "beer"                 y ↦ "Bier"
TASTE(e)                   TRINKEN(e)
THEME(e,x)                 THEME(e,x)
PINT(x)                    GLAS(x)
OF(x,y)                    RELATION(x,y)
BEER(y)                    BIER(y)

           ↘            ↙

           e x y
           TASTE≈TRINKEN(e)
           THEME≡THEME(e,x)
           PINT≈GLAS(x)
           OF⊏RELATION(x,y)
           BEER≡BIER(y)
```

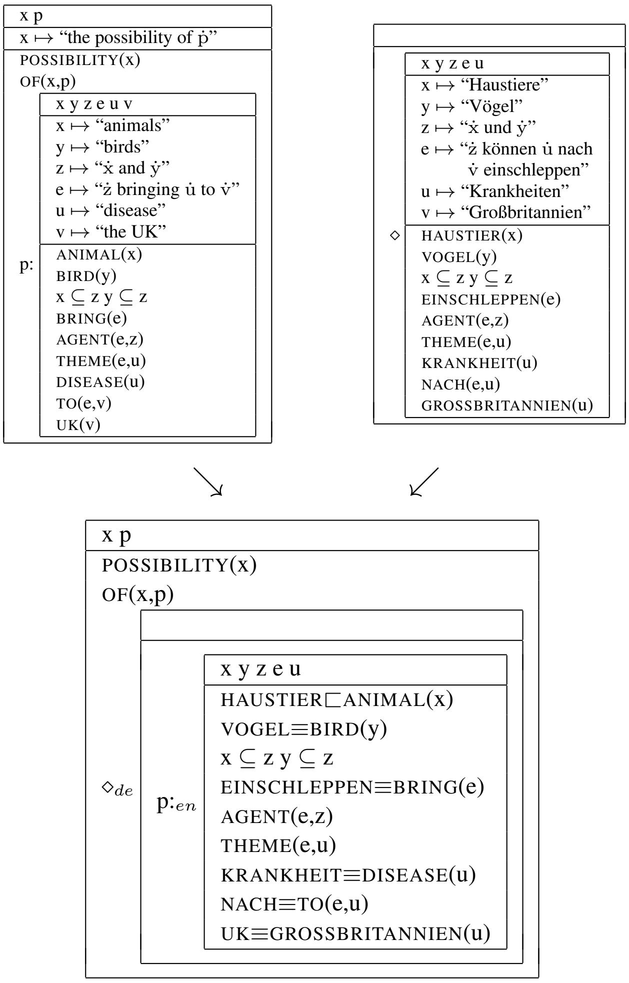
The ≈ combiner is used to align non-logical symbols that have approximately the same meaning, and therefore cannot be described by ≡ or ⊏. It is defined as follows:

$$[S_i{\approx}S_j(x_1,\ldots,x_n)]^{fol} \quad = \quad S_i(x_1,\ldots,x_n) \quad \wedge \quad \forall u_1,\ldots,u_n(\neg S_i(u_1,\ldots,u_n) \rightarrow S_j(u_1,\ldots,u_n))$$

5. Aligning Embedded Contexts

So far we have looked at what we believe are the basic ways to align meaning representations for parallel texts. But there are further issues in meaning alignment, and as a matter of fact the machinery proposed so far isn't able to account for some problems that we encounter when we consider modals and negation. Consider the English sentence "The possibility of animals and birds bringing disease to the UK" and its German translation "Haustiere und Vögel können Krankheiten nach Großbritannien einschleppen." Both sentences contain a modal expression, expressed by a noun in English, and by a modal verb in German. Analogously to Example 1, we could analyze the English modal by introducing a hybrid modal operator (Bos, 2004). Now suppose that the German modal verb is semantically interpreted by the modal possibility operator ◇. This would give the meaning representation as shown in Example 5.

EXAMPLE 5



```
x p                                     
x ↦ "the possibility of ṗ"              
POSSIBILITY(x)                          x y z e u
OF(x,p)                                 x ↦ "Haustiere"
    x y z e u v                         y ↦ "Vögel"
    x ↦ "animals"                       z ↦ "ẋ und ẏ"
    y ↦ "birds"                         e ↦ "ż können u̇ nach
    z ↦ "ẋ and ẏ"                            v̇ einschleppen"
    e ↦ "ż bringing u̇ to v̇"             u ↦ "Krankheiten"
    u ↦ "disease"                       v ↦ "Großbritannien"
    v ↦ "the UK"                     ◇  HAUSTIER(x)
p:  ANIMAL(x)                           VOGEL(y)
    BIRD(y)                             x ⊆ z y ⊆ z
    x ⊆ z y ⊆ z                         EINSCHLEPPEN(e)
    BRING(e)                            AGENT(e,z)
    AGENT(e,z)                          THEME(e,u)
    THEME(e,u)                          KRANKHEIT(u)
    DISEASE(u)                          NACH(e,u)
    TO(e,v)                             GROSSBRITANNIEN(u)
    UK(v)

              ↘                      ↙

x p
POSSIBILITY(x)
OF(x,p)

               x y z e u
               HAUSTIER⊏ANIMAL(x)
               VOGEL≡BIRD(y)
               x ⊆ z y ⊆ z
◇de     p:en   EINSCHLEPPEN≡BRING(e)
               AGENT(e,z)
               THEME(e,u)
               KRANKHEIT≡DISEASE(u)
               NACH≡TO(e,u)
               UK≡GROSSBRITANNIEN(u)
```

There is some discrepancy between the monolingual semantic analyses: the hybrid modal operator (the colon :) that connects a propositional discourse referent with an embedded context in the English case, and the modal operator ◇ in the German case. We could say that in such a case we would need to revise the semantics analysis either on the English or on the German side, to arrive at the same logical operator. An alternative solution, shown here in Example 5, is to decorate logical operators with a *language mode*. This way, we can combine several operators triggered by different languages into one and the same parallel meaning representation. A similar semantic mismatch arises with translating "not recommended" with the German verb "abraten". On the one side we face an explicit negation, and on the other side an implicit negation. Further empirical study is required to shed more light on this issue and evaluate the various possibilities for semantic alignment.

6. Discussion

In this paper we proposed a new formalism to align meaning representations of translated texts. We illustrated the formalism with several non-trivial examples for English–German translations. Certainly, there are many things that we did not consider: light verbs, tense, aspect, discourse relations, pronouns, anaphoric phenomena. Hence, a sensible question to ask is how representative the examples considered in this pilot study are and how and whether this method scales up to other phenomena and languages more distant from English than German.

The only answer we can give to this question is that one just needs to try and investigate, using the empirical method explored here. It is probably fair to point though that the examples that we discussed were not selected because they were easy to model. In fact we tried deliberately to find challenging examples with syntactic mismatches (such as the implicit vs. explicit negation). It seems that for closely related languages such as English and German the approach put forward in this paper is promising. For more distant languages, it could be that the same message is conveyed with very different syntactic structures, as the English–Korean pair[4] ("I have a headache" and its translation "nan-nun meri-ka aphuta") in Example 6.

EXAMPLE 6

$$\begin{array}{|l|}\hline \text{x y e}\\\hline \text{x} \mapsto \text{“I”}\\ \text{y} \mapsto \text{“head”}\\ \text{e} \mapsto \dot{\text{x}}\text{ have a }\dot{\text{y}}\text{ache”}\\\hline \text{HAVE-ACHE(e)}\\ \text{RECIPIENT(e,x)}\\ \text{THEME(e,y)}\\ \text{HEAD(y)}\\\hline\end{array} \qquad \begin{array}{|l|}\hline \text{x y e}\\\hline \text{x} \mapsto \text{“nan”}\\ \text{y} \mapsto \text{“meri”}\\ \text{e} \mapsto \text{“}\dot{\text{x}}\text{-nun }\dot{\text{y}}\text{-ka aphuta”}\\\hline \text{APHUTA(e)}\\ \text{NUN(e,x)}\\ \text{KA(e,y)}\\ \text{MERI(y)}\\\hline\end{array}$$

$$\searrow \qquad\qquad \swarrow$$

$$\begin{array}{|l|}\hline \text{x y e}\\\hline \text{HAVE-ACHE}\approx\text{APHUTA(e)}\\ \text{RECIPIENT}\equiv\text{NUN(e,x)}\\ \text{THEME}\equiv\text{KA(e,y)}\\ \text{HEAD}\equiv\text{MERI(y)}\\\hline\end{array}$$

This is an interesting example because to ensure a smooth alignment between the English and Korean sentence, it forces us to produce a non-literal semantic analysis of the English sentence. It also shows that thematic roles, at least under the analysis put forward here, are more commonly overtly expressed in languages other than English. But then, even within a single language, paraphrases with different syntactic structure should receive similar meaning representations: consider for instance "my head hurts" and "I have a headache". In this particular case, a proper analysis of light verbs would strengthen semantic alignment.

Finally, we would like to remark that the assumptions that we have made for semantic representations are humble: meaning is described with the help of variables, n-place relations, a stock of non-logical symbols, and a couple of logical operators (the usual suspects, i.e. negation, disjunction, modalities). This is standard practice carried out by formal semanticists studying Germanic languages, and we don't see any reason why it wouldn't extend to more distant languages. It is an exercise that could lead not only to interesting language resources for machine translation applications, but also to get a better general understanding of cross-lingual semantic analysis.

7. References

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

Bos, J. (2004). Computational Semantics in Discourse: Underspecification, Resolution, and Inference. *Journal of Logic, Language and Information*, 13(2):139–157.

Fellbaum, C., editor. (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

[4] This example was kindly suggested to me by one of the anonymous reviewers of this paper.