Source-Path-Goal: Investigating the Cross-Linguistic Potential of Frame-Semantic Text Analysis

Source-Path-Goal: Das multilinguale Potenzial framesemantischer Textanalyse

Oliver Čulo, Gerard de Melo, International Computer Science Institute, Berkeley, CA, USA

Summary The ability to analyse natural language semantically has been a long-standing goal in information technology. In the last few decades, the increased availability of annotated text collections and new statistical methods have led to great advances in this area. Frame-semantic annotation, as exemplified by the FrameNet project, describes the meaning of a sentence in terms of the involved cognitive scenarios and the roles of individual participants (e. g. the buyer and the seller in a commerce scenario, but also abstract entities like source, path, and goal in a travel scenario). This contribution studies the cross-linguistic potential of FrameNet from multiple perspectives. It investigates a statistical method to determine frame-semantic relatedness of individual words across languages. It also shows how commonalities and differences between languages in expressing states of affairs can be analysed and extracted from annotated parallel corpora. ►►► **Zusammenfassung** Die Bedeutung sprachlicher Äußerungen analysieren zu können ist schon lange ein großes Ziel der Informationstechnologie. Die zunehmende Verfügbarkeit annotierter Textsammlungen sowie neueste statistische Methoden haben die Wissenschaft diesem Ziel in den letzten Jahrzehnten ein ganzes Stück näher gebracht. Framesemantische Annotation, wie sie im FrameNet-Projekt angewandt wird, beschreibt die Bedeutung eines Satzes mit der damit verbundenen kognitiven Szene sowie den darin vorkommenden Rollen (z. B. der Käufer und der Verkäufer im Handelsszenario, aber auch abstrakte Einheiten wie Source-Path-Goal im Reiseszenario). In diesem Beitrag wird das mehrsprachige Potenzial von FrameNet diskutiert. Es wird eine statistische Methode vorgestellt, mit deren Hilfe die frame-semantische Ähnlichkeit zwischen Wörtern verschiedener Sprachen berechnet werden kann. Außerdem wird gezeigt, inwiefern sich Sprachen im Ausdruck von Zuständen und Ereignissen unterscheiden und wie diese unterschiedlichen Ausdrucksweisen aus parallelen Korpora extrahiert werden können.

Keywords I.2.7 [Computing Methodologies: Artificial Intelligence: Natural Language Processing]; frame semantics, FrameNet, cross-lingual analysis ►►► **Schlagwörter** Frame-Semantik, FrameNet, multilinguale Analyse

1 Introduction

Given the ubiquity of text on the Web, in digital archives, and in human-computer interaction, the ability to analyse natural language semantically has been a long-standing goal in information technology.

In the last few decades, the increased availability of resources like annotated text collections and lexicons combined with new statistical and hybrid methods have led to great advances in this area. Language technology such as spell and grammar checking, parsing, machine translation, speech recognition and question answering are nowadays performing at levels that allow for a number of successful commercial applications, Apple's Siri being one of the most recent prominent examples.

An important trend has been to go beyond the mere surface forms and attempt to uncover more of the underlying concepts and their relations in a given scenario. For instance, a system that detects that "London" in the sentence "In 1897, London sailed north from the Bay Area to join the Klondike Gold Rush" refers to a person (in this case Jack London), not the city of London, can avoid translating it to "Londres" in French. If it can detect that the word "north" here refers to a direction, it will be able to avoid mistranslating that part to German as "segelte London Norden" when it should be "nach Norden" or "nordwärts".

Frame-semantic annotation describes the meaning of a sentence in terms of the involved cognitive scenarios and the roles of individual participants, e. g., the theme (the person who is riding the vessel) as well as the source, path, goal and time in the example sentence above. A frame semantic analysis also reveals for example that the sentence "John sold a car to Mary" essentially describes the same basic situation (*semantic frame*) as "Mary bought a car from John", just from a different perspective. Frame semantic annotation is best exemplified in the FrameNet project [1], an annotated corpus and lexicon that led to the establishment of the task of shallow semantic parsing or automatic semantic role labelling (SRL) in natural language processing and has also been applied to question answering, information extraction and recognizing textual entailment. Overall, FrameNet has been cited over 1500 times.

This contribution studies the cross-linguistic potential of FrameNet and frame-semantic annotation in general, from at least two perspectives. First of all, it investigates statistical methods to determine frame-semantic relatedness of individual words across languages, thereby projecting frame information from English to traditionally lesser-resource languages (i. e. in fact most of the languages in the world).

Second, it shows how commonalities and differences between languages in expressing states of affairs can be analysed and extracted from annotated parallel corpora. From a theoretical perspective, this provides important insights into differences in linguistic strategies and underlying concepts between languages, i. e. data that many disciplines can benefit from. Additionally, such information can also be fed into machine translation systems. Today's MT systems still require major corrections by human post-editors, rendering MT systems unfit for use in many settings.

2 Frame Semantics

Frame semantics [8] is an influential theory of meaning from cognitive linguistics developed by Charles J. Fillmore. The underlying idea of frame semantics is that our knowledge is organised in frames, i. e. representations of prototypical scenes with participants having certain roles in it. These frames are a means of categorizing things we perceive, and allow linguistic and cognitive variations with relation to different perspectives on a scene. They also form the basis for metaphorical uses of language. For a given frame, the FrameNet project [1] provides annotated corpus sentences as well as a frame definition, including a description of the involved participants (*frame elements*, FEs), associated words (*lexical units*, LUs), and their relationships.

An illustrative example is that of the `Commerce_goods-transfer` frame. It usually involves a seller and a buyer, goods transferred, and an amount of money paid for this. However, depending on the perspective, certain roles may or may not be instantiated and lexical realization of the frame evoking element (marked in italics in the following examples) may vary. These perspectives are captured by subframes. With a focus on the price for a transaction, we would produce linguistic instances of the `Commerce_pay` frame, e. g., "$[\text{Vic}]_{\text{Buyer}}$ *paid* $[\$15]_{\text{Money}}$ $[\text{for a ticket}]_{\text{Goods}}$". When focussing on our friend Sam, though, who wanted to get rid of his car, we would, e. g., come up with an instance of the `Commerce_sell` frame like "$[\text{Sam}]_{\text{Seller}}$ *sold* $[\text{the car}]_{\text{Goods}}$ $[\text{to Tony}]_{\text{Buyer}}$".

3 Translating Lexical Units

A first step in making FrameNet multilingual is connecting words in many languages to the existing English LUs and frames using translation techniques. This is an important goal because it allows us to translate individual words using the right sense if we have a frame-semantic analysis of the source sentence. For example, the word "drove" in "She drove the sheep off the road" needs to be translated as "treiben" in German, not as "fahren", which refers to operating a vehicle.

Perhaps more importantly, connecting words to FrameNet's LUs also allows us to project the many years of work that have gone into FrameNet's lexicon to other languages to the extent possible. The largest FrameNet database currently exists for English, with smaller adaptations having been created for languages like Spanish, German, Brazilian Portuguese. Automatically projecting the lexicon of FrameNet using statistical methods constitutes a first major step towards obtaining a full-fledged multilingual FrameNet.

Linking non-English words to English LUs in FrameNet is non-trival because the English translations of a word can be involved in multiple incompatible lexical units if the word has multiple meanings. Each LU requires different non-English words to be attached. As we saw above, the German word "fahren" translates to "drive" in English, but it applies only to the `Operate_vehicle.drive.v` LU, not necessarily, e. g., to the `Cause_motion.drive.v` LU or to the `Subjective_influence.drive.v` LU (as in "They are driven by their political agenda."). Similarly, the French word "épicé" can be translated as "hot" in English, but only with respect to the `Chemical_`

`sense_description.hot.a` LU (which refers to spiciness), not to the `Temperature.hot.a` LU.

3.1 Supervised Disambiguation

We solve this problem by learning to assess possible links based on graph-theoretic properties, as depicted in Fig. 1. We first take a non-English word w, look up all of its translations $w' \in \Gamma(w)$ in a translation dictionary graph using a function Γ, and then look up all LUs

$$l \in \bigcup_{w' \in \Gamma(w)} \Gamma'(w')$$

in FrameNet (assuming Γ' returns the set of lexical units associated with a term in FrameNet). This union constitutes the set of all potential link targets that w can be linked to.

We then compute a series of features $f_i(w, l)$ that provide graph-based statistics about the connection strength between the new word w and any FrameNet LU l in this set. For example, in Fig. 1, some features might assess the connection from "épicé" to the various LUs of "hot" to be rather weak, because the word is ambiguous and there are four possible LUs listed in FrameNet. However, we can also compute additional features that take into account the existence of further indirect paths and hence reveal that `Chemical_sense_description.hot.a`. is the best choice: "épicé" also translates to "spicy", which is connected to the same frame. All of the computed features $f_i(w, l)$ are combined to form a vector $\boldsymbol{v}(w, l) = (f_0(w, l), \ldots, f_k(w, l))$. Given labelled feature vectors computed for manually labelled training examples, one can then predict confidence scores $c(w, l)$ for other potential links (w, l) between non-English words and FrameNet LUs and accept only those that are deemed reliable.

3.2 Results

We conducted initial experiments using FrameNet 1.5, which contains 11 829 LUs (8473 distinct words) and 1019 frames. Translation entries came from the collaborative platform Wiktionary and other translation dictionaries and thesauri. As a training set, we selected 779 random links from German terms to candidate LUs. 184 links were evaluated as positive (correct), 595 as incorrect. We used standard linear-kernel SVMs as implemented in LIBSVM [4].

On an independent test set consisting of 366 annotated links created according to the same principles as the training set, we found that the predicted links have a precision of 86.2% at 52.1% recall. Some of the mistakes we noted involved distinctions between causative and inchoative ("I broke the window" vs. "The window broke"), which are modelled in separate frames in FrameNet but are very hard to discriminate automatically when looking at word translations without additional background information.

Overall, the resulting Multilingual FrameNet Index contains over half a million (535 648) links from words in many different languages to FrameNet LUs. For many languages, there are more links than original English LUs in FrameNet. The number of distinct terms when disregarding word senses and part-of-speech is 301 777. There are 49 languages with at least 2000 distinct terms, and several hundred languages overall.

We are currently investigating methods to improve the quality of the disambiguation. An approach that produces high-quality mappings is to connect FrameNet LUs to WordNet [7], a well-known lexical database. If this linking can be performed reliably, then the many non-English words that have already been connected to WordNet [5] can improve the quality of our Multilingual FrameNet Index.

3.3 Implications

These results allow us to produce a multilingual index of FrameNet LUs that can be used to determine which concepts and frames a sentence evokes. Such a resource can be used in various monolingual tasks like word sense disambiguation and recognizing textual entailment [3] for many different languages that lack a dedicated manually created FrameNet version.

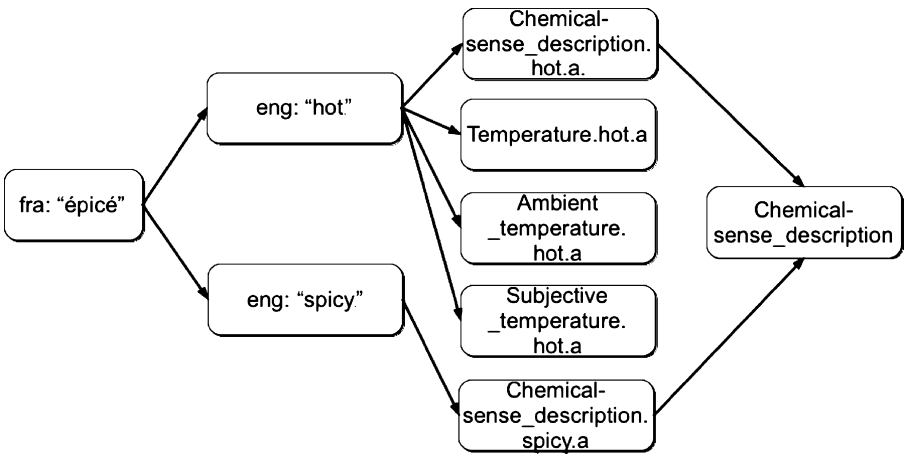

Figure 1 Connections between non-English words and FrameNet's English LUs.

Additionally, the resource can be used cross-linguistically. Given two sentences in two different languages, we can now use the Multilingual FrameNet Index to automatically assess whether they are cognitively related in terms of the frames that they evoke. For instance, the English sentence "He squandered his money" and the German sentence "Er lebte nicht besonders sparsam" can be found to be related because they both evoke the `Frugality` frame, although "squander" and "sparsam" are not translations of each other.

Finally, the fine-grained nature of the Multilingual FrameNet Index can also be exploited to automatically translate individual words from one language to another while taking the specific senses of a word into consideration (e. g., "drive" as "treiben" rather than "fahren" in German).

4 Capturing Cross-Lingual Variation of Framing Strategies

Of course, even beyond the level of individual word choices, languages differ significantly in the way they express one and the same semantic content. For instance, some languages overtly express aspect, while others do not, and those that express aspect may do so in very different ways: Slavic languages distinguish ongoing vs. finished events by lexical means (e. g., the Croatian "popiti" ("drink") will be used after having finished a glass, while its variant "piti" will be used when still drinking), whereas others use grammatical means (e. g., the English progressive forms as in "I was having a beer when I received the call"). Also, subcategorization frames may differ significantly between languages. Although the English word "to walk" corresponds to "gehen" in German, we cannot translate the sentence "Kim walked Pat to the door" using "gehen", because in this case "gehen" cannot accommodate all relevant frame elements. To include "Pat" in the translation, one needs to choose a different verb like "begleiten" ("accompany").

Thus, languages *diverge* in the way they reflect states of affairs. In some cases, these divergences can be explained by cultural differences or by means of analysing the history of a language (e. g., if a language picks up a feature through contact with another language), but in many cases the reasons for (not) having a certain grammatical or lexical feature remain obscure.

Studying differences as well as similarities between languages has nevertheless been useful in many ways, for didactic, translational, technological and other purposes. The groups working on non-English FrameNets have already pointed out various issues when trying to transfer English Frames into their language; many of these issues are well-known and have been described priorly in various translational and cognitive cross-linguistic studies (see Sect. 4.1).

The following paragraphs will discuss some typical divergences that have been described so far, and present a strategy to learn more about such divergences using linguistic data.

4.1 Examples of Known Divergence Patterns

Studies in contrastive linguistics and translation studies have provided some relevant insights into differences in linguistic strategies applied by different languages. For the language pair English-German, Hawkins [9] and König and Gast [12] extensively analyse systematic differences in lexis and grammar. In terms of semantics, one of the typical differences between English and German is the range of semantic types that can appear in subject position, as exemplified by the sentence pair "This hotel forbids dogs" – "In diesem Hotel sind Hunde verboten" [9]. English is far more liberal when it comes to assigning an agentive role to a non-sentient entity, as is done with "The hotel", where German does not typically allow this (though such formulations can sometimes be found in translations or in journalese). Another in-depth comparative study has been performed by Vinay and Darbelnet [17] for English and French. Vinay & Darbelnet reveal that the *chassé croisé*, the crossover switch, is a frequent means of translating between the two languages, as indicated by the co-indexation in this example: "Blériot traversa$_1$ la Manche en avion$_2$" – "Blériot flew$_2$ across$_1$ the Channel" [17]. Slobin [15] shows that Germanic and Romance languages differ in general in the way they frame motion. As exemplified by the previous sentence pair, Romance languages focus on the direction of a motion in the verb of the sentence ("traversa"), moving the manner of motion to a satellite position as adverb ("par avion"). In general, the opposite is true for Germanic languages.

4.2 Harvesting Divergence Patterns from Parallel Corpora

The broad consensus nowadays seems to be that natural language phenomena and handling best be learnt from "real-world data", i. e. from linguistic corpora. While the disadvantage of this approach is that rare phenomena may not be covered at all by the data, the advantage is that knowledge can be automatically extracted, the manual compilation of which would require tedious, time-consuming labour.

One goal of the current work on multilingual issues in frame semantics is that of harvesting divergence patterns in framing between two languages from frame-semantically and grammatically annotated parallel corpora. A parallel corpus here is understood as a corpus consisting either of original texts and their translations, or of translations from one and the same source into multiple languages.[1]

One such instance of a frame-semantically and grammatically annotated corpus is a corpus of about 1000

[1] Using a corpus consisting of translations only is suboptimal, as texts in all languages will show inferences from the source. However, this shall not be a matter of further discussion in this contribution.

parallel sentences annotated with phrase structure and frames and automatically aligned [14]. This corpus shall serve as basis for extracting rules as to how frames and constructions correspond or differ.

A rule is a set of two structures, one structure for each language with frame and phrase structure information, assumed to be equivalent in a given context. A schematic image of such a structure is given in Fig. 2.

The potential of frame-based translation rules has been investigated by Boas [2]. In Boas' idea of a frame-based lexicon, frames serve as an intermediate representation onto which valency patterns of verbs from different languages can be mapped. The valency patterns for the English motion verb "walk" and its German counterpart "gehen" would be annotated as follows (following examples adapted from [2]):

[Kim]$_{\text{Self-mover:NP.Ext}}$ [walked]$_{\text{Self-motion:Target}}$
[to the store]$_{\text{Goal:PP\_to}}$

[Kim]$_{\text{Self-mover:NP.Ext}}$ [ging]$_{\text{Self-motion:Target}}$
[zum Geschäft]$_{\text{Goal:PP\_zu}}$

This being a simple idea, the potential of frame-based annotation shows in cases more difficult for MT, such as partial overlaps in meaning. In the following example, the sense of "walk" used in English cannot be translated into German by "gehen":

[Kim]$_{\text{Self-mover:NP.Ext}}$ [walked]$_{\text{Cotheme-motion:Target}}$
[Pat]$_{\text{Cotheme:NP.Ext}}$ [to the door]$_{\text{Goal:PP\_to}}$

[Kim]$_{\text{Self-mover:NP.Ext}}$ [begleitete]$_{\text{Cotheme-motion:Target}}$
[Pat]$_{\text{Cotheme:NP.Obj}}$ [zum Geschäft]$_{\text{Goal:PP\_zu}}$

The translation of "walk" with "begleiten" is triggered by the Cotheme which is present only in the second example.

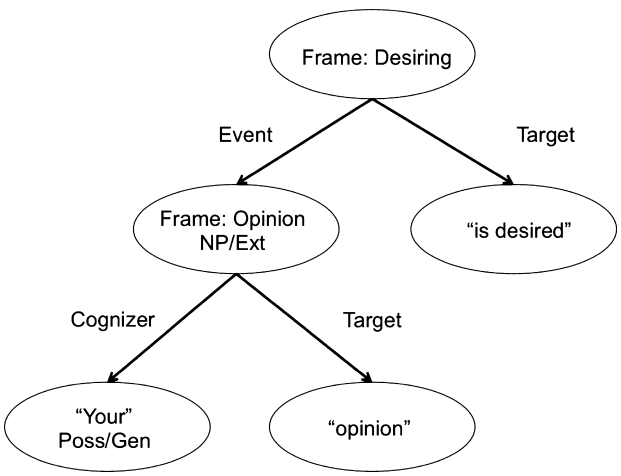

Figure 2 Frames, roles and their syntactic fillers (adapted from [6], with permission from the authors).

Rules extracted from this parallel corpus will be fed into Mutaphrase [6]. Mutaphrase is a system originally designed to create paraphrases departing from a frame semantic input. We will make use of this paraphrasing mechanism in order to perform translation experiments on sentences from English to German and vice versa, given frame semantic and phrase structure annotation.

5 Conclusion

Frame semantics offers a completely new approach to dealing with semantics as opposed to classical formalisms using truth-oriented formal semantics. Amongst other things, frame-semantic annotation may help solve problems like polysemy or partial overlap in meaning in applications like machine translation.

Our findings show that FrameNet indeed has great cross-lingual potential. Frame-semantic annotations have shown to be a valuable resource for exploring conceptual differences between languages, the formalization of which has a positive impact on a number of computational applications involving the analysis of meaning in natural language expressions.

However, there still are significant challenges that need to be addressed. These range from improvements in statistical methods for mapping frame-semantic target words from one language onto another to deeper understanding and more fine-grained formalization of underlying conceptual structures. Our research will continue to evolve in these directions.

References

[1] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet Project. In: *Proc. of of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th Int'l Conf. on Computational Linguistics (COLING-ACL 1998)*, pp. 86–90, 1998.

[2] H. C. Boas. Bilingual FrameNet Dictionaries for Machine Translation. In: *Proccedings of the 3rd Int'l Conf. on Language Resources and Evaluation*, pp. 1364–1371, 2002.

[3] A. Burchardt and A. Frank. Approximating Textual Entailment with LFG and FrameNet Frames. In: *Proc. of the 2nd PASCAL Recognizing Textual Entailment Workshop*, pp. 92–97, 2006.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] G. de Melo and G. Weikum. Towards a universal wordnet by learning from combined evidence. In: *Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM)*, pp. 513–522, 2009.

[6] M. Ellsworth and A. Janin. Mutaphrase: Paraphrasing with FrameNet. In: *Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 143–150, 2007.

[7] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[8] C. J. Fillmore. Frame Semantics and the Nature of Language. In: *Annals of the New York Academy of Sciences* 280, pp. 20–32, 1976.

[9] J. A. Hawkins. *A Comparative Typology of English and German. Unifying the Contrasts*. London: Croom Helm, 1986.

[10] H. Hoang, P. Koehn, and A. Lopez. A Unified Framework for Phrase-based, Hierarchical, and Syntax-based Statistical Machine Translation. In: *Proc. of the Int'l Workshop on Spoken Language Translation*, pp. 152–159, 2009.

[11] P. Koehn, F. J. Och, and D. Marcu. Statistical Phrase-Based Translation. In: *Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003). Proc. of the Conf.*, pp. 127–133, 2003.
[12] E. König, and V. Gast. *Understanding English-German Contrasts*. Berlin: Erich Schmidt Verlag, 2007.
[13] S. Padó and K. Erk. To Cause or Not to Cause: Cross-lingual Semantic Matching for Paraphrase Modelling. In: *Proc. of the Cross-Language Knowledge Induction Workshop*, 2005.
[14] S. Padó and M. Lapata. Cross-lingual Projection of Role-semantic Information. In: *Proc. of the Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pp. 859–866, 2005.
[15] D. I. Slobin. The Many Ways to Search for a Frog: Linguistic Typology and the Expression of Motion Events. In: *Relating Events in Narrative: Typological Perspectives*, edited by S. Strömqvist and L. Verhoeven, pp. 219–257, 2004.
[16] L. Tesnière. *Éléments de Syntaxe Structurale*. Paris: Klinksieck, 1959.
[17] J.-P. Vinay and J. Darbelnet. *Stylistique comparée du français et de l'anglais. Méthode De Translation*. Paris: Didier, 1958.

Received: February 29, 2012, accepted: March 18, 2012

Dr. phil. Oliver Čulo obtained his doctoral degree with distinction as a student of the Chair for Machine Translation at Saarland University, Saarbrücken, in 2011. In 2008, he started a teaching and research position at the Faculty of Translation Studies at Mainz University. His teaching, research and publications have been devoted to topics at the crossroads of contrastive linguistics, translation studies and multilingual natural language processing. He joined the AI-FrameNet group at ICSI in October 2011.

Address: International Computer Science Institute, 1947 Center St., 94704 Berkeley, CA, USA, e-mail: oculo@icsi.berkeley.edu

Dr.-Ing. Gerard de Melo, as a member of the Databases and Information Systems group at the Max Planck Institute for Informatics, Gerard de Melo obtained his doctoral degree with distinction from Saarland University, Saarbrücken in 2010. He has won two Best Paper Awards (ICGL 2008, CIKM 2010) and one Best Demo Award (WWW 2011). Following a 4-month internship at Microsoft Research Cambridge, he joined the Artificial Intelligence group at the ICSI in 2011, working in the FrameNet group.

Address: International Computer Science Institute, 1947 Center St., 94704 Berkeley, CA, USA, e-mail: demelo@icsi.berkeley.edu