Two Approaches to Automatic Matching of Atomic Grammatical Features in LFG

Anton Bryl
CNGL, Dublin City University
abryl@computing.dcu.ie

Josef van Genabith
CNGL, Dublin City University
josef@computing.dcu.ie

The recent increase in attention to Lexical-Functional Grammars in syntax-based statistical machine translation (SMT) [1, 2] poses new problems for processing richly annotated data. Alignment is a core issue in machine translation. In our work we focus on deepening the automatic cross-language structure alignment by adding the possibility to effectively align not only words, but also atomic f-structure features. Though sets of atomic features (such as case, number, etc.) differ for different languages, they are far from being disjoint. A number of features, such as number or case are shared between many languages. It is common practice to hardcode these similarities in the grammars; that is, to give the same names to the same or similar linguistic properties in the grammars for different languages. However, when one wants to make use of correspondences between such features in the framework of a language-agnostic syntax-based SMT system, such "feature name alignment" between source and target grammars cannot simply be taken for granted. Moreover, the degree of correspondence may differ from feature to feature across grammars. This motivates the need for an automatic way to judge correspondences between atomic features in f-structure representations for arbitrary language pairs.

We show that, provided we have parsers for two languages and a parallel corpus for these languages, it is possible to automatically identify at least part of the correspondences between atomic-valued grammatical features. Once identified, these feature pairs can further be used to improve the coverage of transfer-based machine translation. For example, if the algorithm identifies that NUM in English and NUM in German generally co-vary (we presume again, that we do not use any prior knowledge about this correspondence), then we can safely induce transfer rules (from aligned parsed bitext corpora) which abstract over the number feature, providing an effective back-off to more specific transfer rules.

Another application in which automatic grammatical feature matching is potentially useful is parallel grammar design. More specifically, feature matching is able to provide empirical evidence of the similarity between features of different languages, thus helping to determine whether they are to be treated as the same or different features.

We define and evaluate two methods of grammatical feature matching.

Method 1. The idea behind the first method is that if a feature **A** in one language corresponds to feature **B** in another language, then a change in the value of **A** in the source language (SL) frequently corresponds to a change

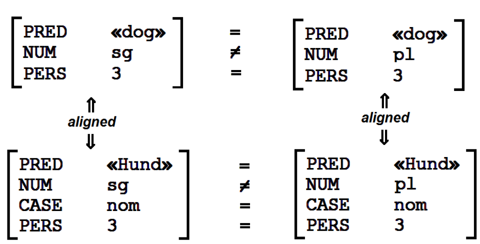

Figure 1: Simultaneous change of the values of `NUM` in parallel data. Finding such a situation, the algorithm increases the probability counter for the (`Eng.NUM` $\Rightarrow$ `Ger.NUM`) correspondence.

German-to-English			English-to-German		
Pair	N_{it}	$\frac{C_{A,B}}{C_A}$	*Pair*	N_{it}	$\frac{C_{A,B}}{C_A}$
(NUM ⇒ NUM)	1	0.88	(NUM ⇒ NUM)	1	0.67
(TNS-ASP ⇒ TNS-ASP)	1	0.70	(TNS-ASP ⇒ TNS-ASP)	1	0.81
(CLAUSE-TYPE ⇒ CLAUSE-TYPE)	1	0.62	(CASE ⇒ CASE)	1	0.86
(CASE ⇒ CASE)	1	0.51	(DEGREE ⇒ DEGREE)	1	0.98
(ATYPE ⇒ ATYPE)	1	0.79	(PASSIVE ⇒ PASSIVE)	2	0.55
(COMP-FORM ⇒ COMP-FORM)	2	0.92			
(PASSIVE ⇒ PASSIVE)	2	0.64			

Table 1: Experimental results for Method 1. N_{it} is the number of iteration on which the pair emerged. $\frac{C_{A,B}}{C_A}$ is the normalized score (see algorithm).

Pair	**Inc. of prediction accuracy**	**Pair**	**Inc. of prediction accuracy**
NUM ⇒ NTYPE	0.17	NUM ⇒ PERS	0.25
NUM ⇒ NUM	0.35	NUM ⇒ CASE	0.15

Table 2: A part of experimental results for Method 2 (German-to-English), showing the increase of prediction accuracy over the pick-most-frequent baseline. Apart from the NUM, which is the correct match, CASE, PERS, NTYPE also gained high scores. This is due to their frequent co-presence with NUM in f-structures, which results in accurate prediction of the *feature absent* special value. However, the correct match clearly outscores them. All the other features got 0 score when matched with NUM, and are not included in the table for the sake of space.

in the value of **B** in the aligned translation. More precisely, the method identifies pairs of sub-f-structures in the SL data, which differ only in one atomic feature value, and then checks the difference in the aligned target language sub-f-structures (see Figure 1). The accumulated data from all such pairs is then used to find the best matches between the features. Complex PRED-less features like TNS-ASP are treated as atomic and considered equal if all their child attributes are equal, and unequal otherwise.

Method 2. The second method makes use of the mutual predictability of features of the two languages. For each possible pair (**Lang1.A**, **Lang2.B**) we calculate the best possible accuracy of prediction of the value of **Lang2.B** in the TL structure by the value of **Lang1.A** in the aligned SL f-structure. The accuracy of pick-most-frequent baseline is then subtracted from this value. The resulting value, that is the increase in prediction accuracy over the baseline, is used as a measure of similarity between the features. The absence of a certain feature in a structure is considered a special *feature absent* value of this feature.

Both methods are evaluated experimentally on 219,667 sentences of parced Europarl [3] German-English data and show promising results. The results for two iterations of Method 1 are presented in Table 1. Table 2 contains some example numbers for Method 2.

References

[1] E. Avramidis and J. Kuhn. Exploiting XLE's finite state interface in LFG-based statistical machine translation. In *Proceedings of LFG'09*, 2009.

[2] Y. Graham and J. van Genabith. An open source rule induction tool for transfer-based SMT. *The Prague Bulletin of Mathematical Linguistics*, Special Issue: Open Source Tools for Machine Translation, 91:37–46, 2009.

[3] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, pages 79–86, 2005.