Learning Tense Translation from Bilingual Corpora

Michael Schiehlen*
Institute for Computational Linguistics, University of Stuttgart,
Azenbergstr. 12, 70174 Stuttgart
mike@adler.ims.uni-stuttgart.de

Abstract

This paper studies and evaluates disambiguation strategies for the translation of tense between German and English, using a bilingual corpus of appointment scheduling dialogues. It describes a scheme to detect complex verb predicates based on verb form subcategorization and grammatical knowledge. The extracted verb and tense information is presented and the role of different context factors is discussed.

1 Introduction

A problem for translation is its context dependence. For every ambiguous word, the part of the context relevant for disambiguation must be identified (disambiguation strategy), and every word potentially occurring in this context must be assigned a bias for the translation decision (disambiguation information). Manual construction of disambiguation components is quite a chore. Fortunately, the task can be (partly) automated if the tables associating words with biases are learned from a corpus. Statistical approaches also support empirical evaluation of different disambiguation strategies.

The paper studies disambiguation strategies for tense translation between German and English. The experiments are based on a corpus of appointment scheduling dialogues counting 150,281 German and 154,773 English word tokens aligned in 16,857 turns. The dialogues were recorded, transcribed and translated in the German national Verbmobil project that aims to develop a tri-lingual spoken language translation system. Tense is interesting, since it occurs in nearly every sentence. Tense can be expressed on the surface lexically as well as morphosyntactically (analytic tenses).

* This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01 IV 101 U. Many thanks are due to G. Carroll, M. Emele, U. Heid and the colleagues in Verbmobil.

2 Words Are Not Enough

Often, sentence meaning is not compositional but arises from combinations of words (1).

(1) a. Ich habe ihn gestern gesehen.
I have him yesterday seen
I saw him yesterday.

b. Ich schlage Montag vor.
I beat Monday forward
I suggest Monday.

c. Ich möchte mich beschweren.
I 'd like to myself weigh down
I'd like to make a complaint.

For translation, the discontinuous words must be amalgamated into single semantic items. Single words or pairs of lemma and part of speech tag (L-POS pairs) are not appropriate. To verify this claim, we aligned the L-POS pairs of the Verbmobil corpus using the completely language-independent method of Dagan et al. (1993). Below find the results for *sehen*[1] (*see*) in order of frequency and some frequent alignments for reflexive pronouns.

72	sehen:VVFIN	be:VBZ	(*aussehen*)
44	sehen:VVFIN	do:VBP	(do-support)
39	sehen:VVFIN	have:VBP	(perfect)
35	sehen:VVFIN	see:VB	
176	wir:PRF	meet:VB	(sich treffen)
33	wir:PRF	we:PP	
30	sich:PRF	spell:VBN	(sich schreiben)
16	ich:PRF	forward:RP	(sich freuen auf)
14	wir:PRF	agree:VB	(sich einigen)
13	ich:PRF	myself:PP	

[1] The prefix verb *aus-sehen* (*look, be*) is very frequent in the corpus, it often occurs in questions. Present *sehen* was frequently translated into perfect *discover*.

3 Partial Parsing

A full syntactic analysis of the sort of unrestricted spoken language text found in the Verbmobil corpus is still beyond reach. Hence, we took a partial parsing approach.

3.1 Complex Verb Predicates

Both German and English exhibit complex verb predicates (CVPs), see (2). Every verb and verb particle belongs to such a CVP and there is only one CVP per clause.

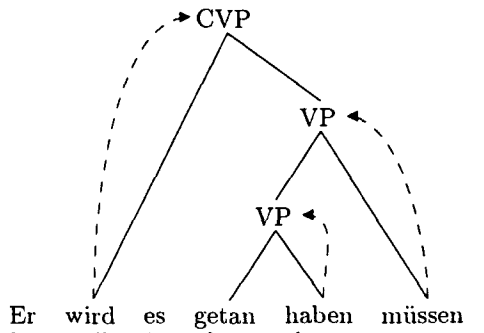
(2) He *would* not *have called* me *up*.

The following two grammar fragments describe the relevant CVP syntax for English and German. Every auxiliary verb governs only one verb, so the CVP grammar is basically[2] regular and implementable with finite-state devices.

$S \rightarrow \ldots VP \ldots$
$VP \rightarrow$ **hd**:V (to) VP
$VP \rightarrow$ **hd**:V ... (Particle)

$S \rightarrow \ldots$ **hd**:$V_{fin} \ldots$ (Refl) ... VC ...
$S \rightarrow \ldots$ (Refl) ... VC ...
$S \rightarrow \ldots$ VC **hd**:$V_{fin} \ldots$ (Refl) ...
$VC \rightarrow$ (VC) (zu) **hd**:V
$VC \rightarrow$ SeparatedVerbPrefix

English CVPs are left-headed, while German CVPs are partly left-, partly right-headed.



He will have to have done it.

[2]The grammar does not handle insertion of CVPs into other CVPs and partially fronted verb complexes (3).

(3) Versuchen hätte ich es schon gerne wollen.
try 'd have I it liked to
I'd have liked to try it.

3.2 Verb Form Subcategorization

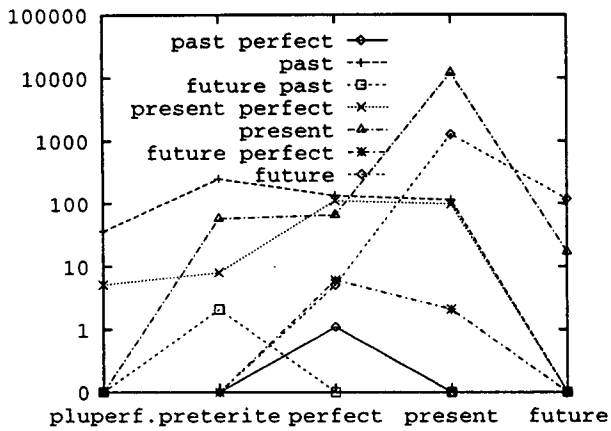
Auxiliary verbs form a closed class. Thus, the set $sub(v)$ of infinite verb forms for which an auxiliary verb v subcategorizes can be specified by hand. English and German auxiliary verbs govern the following verb forms.

- infinitive e.g. *will*
- *to*-infinitive (T) e.g. *want*
- past participle (P) e.g. *get*
- P ∨ T e.g. *have*
- present participle ∨ P ∨ T e.g. *be*

- infinitive (I) e.g. *müssen*
- *zu*-infinitive (Z) e.g. *scheinen*
- perf.part. with *haben* (H) e.g. *bekommen*
- H ∨ I e.g. *werden*
- H ∨ I ∨ Z e.g. *haben*
- perf.part. with *sein* ∨ H ∨ I ∨ Z e.g.*sein*

3.3 Transducers

Two partial parsers (rather: transducers) are used to detect English and German CVPs and to translate them into predicate argument structures (*verb chains*). The parsers presuppose POS tagging and lemmatization. A data base associates verbs v with sets $mor(v)$ of possible tenses or infinite verb forms.
Let $m = |\{mor(v) : \text{Verb } v\}|$ and $n = |\{sub(v) : \text{Verb } v\}|$. Then the English CVP parser needs $n + 1$ states to encode which verb forms, if any, are expected by a preceding auxiliary verb. Verb particles are attached to the preceding verb. The German CVP parser is more complicated, but also more restrictive as all verbs in a verb complex (VC) must be adjacent. It operates in left-headed (S) or right-headed mode (VC). In VC-mode (i.e. inside VCs) the order of the verbs put on the output tape is reversed. In S-mode, $n + 1$ states again record the verb form expected by a preceding finite verb V_{fin}. VC-mode is entered when an infinite verb form is encountered. A state in VC-mode records the verb form expected by V_{fin} $(n+1)$, the infinite verb form of the last verb encountered (m), and the verb form expected by the VC verb, if the VC consists of only one verb $(n+1)$. So there are $m*(n+1)^2$ states. As soon as a non-verb is encountered in VC-mode or the verb form of the previous verb does not fit the subcategorization requirements of the current verb, a test is performed to see if the verb form of the last verb



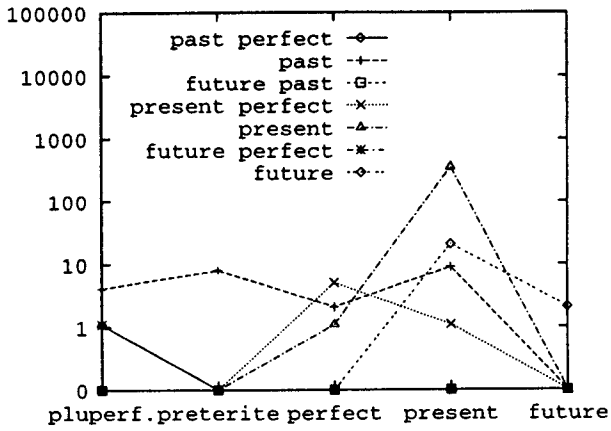


Figure 1: translation frequencies G→E (left: simple tenses, right: progressive tenses)

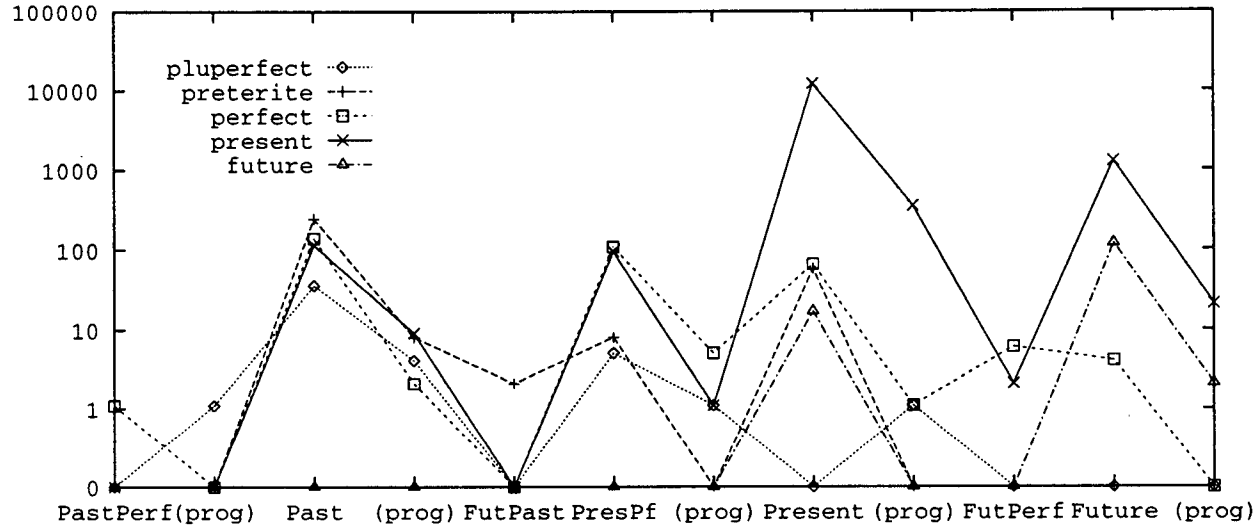


Figure 2: translation frequencies E→G

in VC fits the verb form required by V_{fin}. If it does or there is no such finite verb, one CVP has been detected. Else V_{fin} forms a separate CVP. In case the VC consists of only one verb that can be interpreted as finite, the expected verb form is recorded in a new S-mode state. Separated verb prefixes are attached to the finite verb, first in the chain.

3.4 Alignment

In the CVP alignment, only 78 % of the turns proved to have CVPs on both sides, only 19 % had more than one CVP on some side. CVPs were further aligned by maximizing the translation probability of the full verbs (yielding 16,575 CVP pairs). To ensure correctness, turns with multiple CVPs were inspected by hand. In word alignment inside CVPs, surplus tense-bearing auxiliary verbs were aligned with a tense-marked NULL auxiliary (similar to the English auxiliary *do*).

3.5 Alignment Results

The domain biases the corpus towards the future. So only 5 out of 6 German tenses and 12 out of 16 English tenses occurred in the corpus. Both *will* and *be going to* were analysed as future, while *would* was taken to indicate conditional mood, hence present.

- present (15,710)
- preterite (331)
- future (150)
- perfect (344)
- pluperfect (49)

- present (12,252; progressive: 358)
- past (594; progressive: 23)
- present perfect (227; progressive: 7)
- past perfect (1; progressive: 1)
- future (1,429; progressive: 23)
- future perfect (10) • future in the past (3)

In some cases, tense was ambiguous when considered in isolation, and had to be resolved in tandem with tense translation. Ambiguous tenses on the target side were disambiguated to fit the particular disambiguation strategy.

- G present/perfect (*verreist sein*) (39)
- G present/past (*sollte, ging*) (229)
- E pres./present perfect (*have got*) (500)
- E pres./past (*should, could, must*) (1,218)

4 Evaluation

Formally, we define source tense and target tense as two random variables S and T. Disambiguation strategies are modeled as functions tr from source to target tense. *Precision* figures give the proportion of source tense tokens t_s that the strategy correctly translates to target tense t_t, *recall* gives the proportion of source-target tense pairs that the strategy finds out.

$$(4)\quad precision_{tr}(t_s, t_t) = P(T = t_t | S = t_s, tr(t_s) = t_t)$$
$$recall_{tr}(t_s, t_t) = P(tr(t_s) = t_t | S = t_s, T = t_t)$$

Combined precision and recall values are formed by taking the sum of the frequencies in numerator and denominator for all source and target tenses. Performance was cross-validated with test sets of 10 % of all CVP pairs.

4.1 Baseline

A baseline strategy assigns to every source tense the most likely target tense ($tr(t_s) = \arg\max_{t_t} P(t_t|t_s)$, strategy t). The most likely target tenses can be read off Figures 1 and 2. Past tenses rarely denote logical past, as discussion circles around a future meeting event, they are rather used for politeness.

(5) a. Ich wollte Sie fragen, wie das aussieht.
I wanted to ask you what is on.

b. übermorgen war ich ja auf diesem Kongreß in Zürich.
the day after tomorrow, I'll be (lit: was) at this conference in Zurich.

4.2 Full Verb Information

Three more disambiguation strategies condition the choice of tense on the full verb in a CVP, viz. the source verb ($tr(t_s, v_s) = \arg\max_{t_t} P(t_t|t_s, v_s)$, strategy v_s), the target verb ($tr(t_s, v_t)$, strategy v_t), and the combination of source and target verb ($tr(t_s, \langle v_s, v_t \rangle)$, strategy v_{st}). The table below gives precision and recall values for these strategies and for the strategies obtained by smoothing (e.g. v_{st}, v_s, v_t, t is v_{st} smoothed first with v_s, then with v_t, and finally with t). Smoothing with t results in identical precision and recall figures.

	G→E			E→G		
	prec.	recall	, t	prec.	recall	, t
t	.865	.865	.865	.957	.957	.957
v_s	.885	.854	.879	.970	.941	.965
v_t	.900	.876	.896	.973	.933	.966
v_{st}	.916	.819	.899	.979	.874	.965
v_{st}, v_t, v_s	.902	.892	.900	.970	.956	.967
v_{st}, v_s, v_t	.899	.889	.897	.971	.957	.967

We see that inclusion of verb information improves performance. Translation pairs approximate the verb semantics better than single source or target verbs. The full verb contexts of tenses can also be used for verb classifications.

Aspectual classification: The aspect of a verb often depends on its reading and thus can be better extrapolated from an aligned corpus (e.g. *I am having a drink (trinken)*). German allows punctual events in the present, English prefers present perfect (e.g. *sehen, finden, feststellen(discover, find, see); einfallen (occur, remember); treffen, erwischen, sehen (meet)*).

World knowledge: In many cases perfect maps an event to its result state.

finish	⇒ fertig sein
forget	⇒ nicht mehr wissen
denken an	⇒ have in mind
sich verabreden	⇒ have an appointment
sich vertun	⇒ be wrong
settle a question	⇒ (the question) is settled

4.3 Subordinating Conjunctions

Conjunctions often engender different mood.

- In conditional clauses English past tenses usually denote present tenses. Interpreting hypothetical past as present increases performance by about 0.3 %.

• In subjunctive environments logical future is expressed by English simple present. The verbs *vorschlagen (suggest)* (in 11 out of 14 cases) and *sagen (say)* (2/5) force simple present on verbs that normally prefer a translation to future.

(6) I suggest that we meet on the tenth.

• Certain matrix verbs[3] trigger translation of German present to English future.

4.4 Representation of Tense

Tense can not only be viewed as a single item (as sketched above, representation r_t). In compositional analyses of tense, source tense S and target tense T are decomposed into components $\langle S_1, \ldots, S_n \rangle$ and $\langle T_1, \ldots, T_n \rangle$. A disambiguation strategy tr is correct if $\forall i : tr(S_i) = T_i$.

One decomposition is suggested by the encoding of tense on the surface ($\langle$present/past, 0/*will/be going to/werden*, 0/*have/haben/sein*, 0/*be*$\rangle$, representation r_s). Another widely used framework in tense analysis (Reichenbach, 1947) ($\langle$E</≈/>R, R</≈/>S, ±progr$\rangle$, representation r_r) analyses English tenses as follows:

	R≈S	R<S	R>S
E≈R	present	past	
E<R	present perf.	past perf.	fut. perf.
E>R	future	future past	

A similar classification can be used for German except that present and perfect are analysed as ambiguous between present and future (E≥R≈S and E<R≥S).

repr.	strat.	G→E prec.	G→E recall	G→E $,t$	E→G prec.	E→G recall	E→G $,t$
r_t	t	.865	.865	.865	.957	.957	.957
r_s	t	.859	.859	.859	.955	.955	.955
r_s	v_s	.883	.853	.876	.966	.938	.961
r_s	v_t	.894	.871	.890	.971	.933	.964
r_s	v_{st}	.912	.815	.894	.978	.874	.962
r_r	t	.861	.861	.861	.964	.964	.964
r_r	v_s	.885	.855	.879	.973	.945	.970
r_r	v_t	.898	.875	.894	.977	.939	.972
r_r	v_{st}	.915	.817	.897	.982	.878	.970

The poor performance of strategy r_s corroborates the expectation that tense disambiguation is helped by recognition of analytic tenses. Strategy r_r performs slightly worse than r_t. The really hard step with Reichenbach seems to be the mapping from surface tense to abstract representation (e.g. deciding if (polite) past is mapped to logical present or past). r_r performs slightly better in E→G, since the burden of choosing surface tense is shifted to generation.

repr.	strat.	G→E prec.	G→E recall	G→E $,t$	E→G prec.	E→G recall	E→G $,t$
$r_{r'}$	t	.861	.861	.861	.957	.957	.957
$r_{r'}$	v_s	.883	.853	.877	.968	.940	.963
$r_{r'}$	v_t	.895	.872	.891	.971	.933	.965
$r_{r'}$	v_{st}	.913	.816	.895	.979	.875	.964

5 Conclusion

The paper presents a way to test disambiguation strategies on real data and to measure the influence of diverse factors ranging from sentence internal context to the choice of representation. The pertaining disambiguation information learned from the corpus is put into action in the symbolic transfer component of the Verbmobil system (Dorna and Emele, 1996).

The only other empirical study of tense translation (Santos, 1994) I am aware of was conducted on a manually annotated Portuguese–English corpus (48,607 English, 43,492 Portuguese word tokens and 6,334 tense translation pairs). It neither gives results for all tenses nor considers disambiguation factors. Still, it acknowledges the surprising divergence of tense across languages and argues against the widely held belief that surface tenses can be mapped directly into an interlingual representation. Although the findings reported here support this conclusion, it should be noted that a bilingual corpus can only give one of several possible translations.

[3] *ausgehen von, denken, meinen (think), hoffen (hope), schade sein (be a pity)*

References

Ido Dagan, Kenneth W. Church, and William A. Gale. 1993. Robust Bilingual Word Alignment for Machine-Aided Translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1–8.

Michael Dorna and Martin C. Emele. 1996. Semantic-Based Transfer. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, Copenhagen, Denmark.

Hans Reichenbach. 1947. *Elements of Symbolic Logic*. Macmillan, London.

Diana Santos. 1994. Bilingual Alignment and Tense. In *Proceedings of the Second Annual Workshop on Very Large Corpora*, pages 129–141, Kyoto, August.