TRANSLATION RELATIONS AND THE COMBINATION OF ANALYTICAL AND STATISTICAL METHODS IN MACHINE TRANSLATION

Hubert Lehmann, Nikolaus Ott
IBM Deutschland GmbH, Scientific Center,
Institute for Knowledge Based Systems
Wilckensstr. 1a, D-6900 Heidelberg, Germany.
Email: LEH at DHDIBM1.BITNET

Abstract

One of the determining factors in translation relations is semantic compatibility of cooccurring words. Methods for discovering and representing semantic compatibility relations and their role in translation relations are discussed. The methods used are a combination of analytical and statistical methods, and they are all based on an analytical model of translation. Using the case of semantic compatibility as an example, it is argued that in order to achieve progress in linguistics in general and in Machine Translation in particular neither a purely empiricist nor a purely rationalist approach will do, but that a combination of models and methods is required.

Introduction

The key problem of the theory of Machine Translation (MT) can be formulated as the question whether an algorithm can be devised which translates as well from one language to another as an ideal translator would. This implies that the theory of MT and its components - among them a theory of verbal communication and a theory of comparing languages — must be formal theories. That such theories, if they exist, can be simply derived from observed data is an assumption apparently made by some who use statistical methods in linguistics and in particular in MT, and violently denied by Quine (1953) as well as Chomsky (1986). We advocate a pragmatist approach (as in Lehmann, 1973) which allows for abstract theoretical constructs, but we insist that they must have consequences which can be tested. We prefer operational definitions which can be implemented in algorithms used to analyze and generate texts.

A key question of the theory of translation is how can we decide whether an utterance in one language is a translation of an utterance in another language, or, since we should admit that translations cannot always be perfect, whether it is a **good** translation of an utterance in another language. According to Quine (1959) translation amounts to a **semantic correlation** of sentences. In a realistic theory of translation we must be interested in larger as well as smaller linguistic

units than sentences, and we must consider pragmatic as well as semantic correlations. We are interested in the same **communicative effect** on the respective hearers of an utterance and its translation.

A key assumption of the theory of translation is that translations can be produced **compositionally,** where compositionality should be understood in a relatively broad way which allows to include contextual parameters determined outside the boundaries of a sentence.

The compositionality assumption underlies the compilation of lists of translation equivalents for single words or short expressions as it is done in conventional dictionaries. It also underlies the work on semantic compatibility which is reported on here and which is being undertaken in the framework of the TransLexis project at the IBM Germany Scientific Center.

The TransLexis project aims at an empirically and theoretically based description of lexical information and is to provide the lexicon technology for the LMT (Logic-programming based Machine Translation) system. Its lexical knowledge base is implemented using the relational database system SQL/DS and Prolog, and it provides a lexicographic workbench for translators and terminologists. The project is intended to lead to the development of a polytheoretic and multifunctional lexicon. For a description see Bläser et al. (1992).

The LMT project was originally set up by M. McCord and a first version was described in McCord (1989). Meanwhile the underlying grammatical formalism used in analysis was modified taking up ideas from McCord (1980). The new formalism is described in McCord (1991), and new features of LMT are presented in Rimon et al. (1991). The LMT project is now being pursued at the IBM T. J. Watson Research Center and at several IBM Scientific Centers involving translation from English into Arabic, Danish, French, German, Hebrew and Spanish, and translation from Danish, German and Spanish into English.

The LMT approach relies strongly on lexicalist syntax, making much use of syntactic relations, it uses semantic compatibility mainly for disambiguation and for the selection of translations. The translation strategy of LMT is transfer oriented: single and multiple word lexemes are replaced by target expressions according to various kinds of criteria, and transformations are applied to produce correct word order in the target language.

We will describe the contextual factors which influence translation relations and explain our view on homonymy and polysemy to give the background for the discussion of syntactic relations and semantic compatibility and their effect on translation relations. Then we will describe the analytical and statistical methods which can be applied to determine semantic compatibility. We will finally show how analytical and statistical methods can complement each other.

Statistics based approaches to MT tend to minimize the analytical assumptions about individual languages and translation relations. Analytical approaches, in contrast, rely on developing analytical models of language and translation which are as fine grained as possible, and they should try to explain why linguistic phenomena — including statistical behavior — are as they are: We argue that good analytical models are not only pleasing to the linguist but are a necessary prerequisite for statistical investigations.

Contextual factors in translation relations

We present and illustrate a number of contextual factors which have been recognized as relevant for determining translation relations:

Text genre: Many types of discourse situations and texts can be characterized as belonging to a certain **genre.** In the framework of the LMT project we look at the treatment of computer manuals. They can be regarded as *a* subgenre of **instructions for use.** It should be noted that a given manual, brochure, or leaflet which comes with a product even though it may say *instructions for use* on its cover may actually contain texts which belong to several different genres, e.g. conditions for guarantee.

An example for how text genre influences translation relations is the rendering of certain infinitives in German by imperatives in English (or vice versa) which is typical for instructions type texts.

> Richtige Spannung *einstellen.*
> *Set* correct voltage.

But imperatives do occur in German instructions texts as well. So other factors must be involved which determine the choice of the construction.

Author: There is a group of factors which can be attributed to the author of an utterance: **author intentions and attitudes, personal (company) style and vocabulary.** With respect to instructions of use, any author will want to avoid describing a product or its properties in terms which may have a negative connotation, and this must be preserved in a translation. Regarding style a company may generally use one term out of several available synonyms.

Audience: Several factors can be related to the intended audience of a text. In instructions for use, the rendering of English imperatives in German will be different depending on whether the intended audience consists of children or adults.

Instructions may address the general public or experts, and this will influence the use of general vocabulary, terminology, and jargon in instructions texts.

Subject area: The influence of the subject area(s) of a given text on the proper translation (and disambiguation) is often discussed in the literature. As an example consider *Spannung* which can be translated as *tension, voltage, excitement, suspense,* in the field of electricity usually as *voltage,* but in certain contexts also as *tension* (cf. *high tension wires).*

Discourse context: As factors of discourse context we want to mention **deixis,** which occurs in texts when references are made to pictures, figures or tables, and **anaphora.** How anaphoric relations influence translation relations is obvious for the gender of pronouns as in:[1]

> Wenn der *Formularstapel (m)* hoch ist, muß *er (m)* eventuell aus dem Formularfach herausgenommen werden.
> If the *forms stack* is high, you might need to take *it* out of the forms compartment.

But the choice of a target word may be influenced as well. This can be seen in:

> Es ist eine große Ungerechtigkeit, daß Radfahrer, die ihr *Rad* auf der Straße abstellen, mit 80 Mark zur Kasse gebeten werden, weil das *Fahrzeug* eine gefährliche Behinderung darstelle.
> It is very unfair that bicyclists who leave their *bike* out on the street are fined 80 Marks because the *vehicle* is supposedly a dangerous hindrance.

If *Rad* had not been coreferent with *Fahrzeug,* it could also (and more probably) have been translated as *wheel.*

Sentence and clause context: Most approaches to grammar in theoretical as well as in computational linguistics use notions such as government (subcategorization, slot frames, valency, etc.) and semantic compatibility to restrict generative power, to disambiguate expressions, or to find proper translation equivalents (see Lehrberger and Bourbeau, 1988). The effect of semantic compatibility on translation relations will be discussed in more detail below.

---

[1] The following two examples are due to Leass (1991)

Homonymy and polysemy

While it is generally agreed that homonymy and polysemy are useful concepts for describing lexical items, there is much less agreement on how homonyms and senses are to be distinguished for a given lexical item. What is needed are operational definitions which provide unequivocal criteria for making the desired distinctions.

We will not use etymological arguments to distinguish homonyms as is often done in dictionaries, because a given text will hardly ever provide contextual clues to determine a homonym on an etymological basis. Thus we will not assume two homonyms for the English word *ball.* Our notion of homonymy uses:

1. **part of speech:** if different parts of speech can be associated with a word, we have different homonyms, e.g. *book(n)* and book(v);

2. **gender:** if different genders can be associated with a word, we have different homonyms, e.g. German *Gehalt* — masculine (content, concentration) and neuter (salary);

3. **inflection:** if different inflection paradigms can be associated with a word, we have different homonyms, e.g. German *hängen, hing, gehangen* (Engl. *hang* intransitive) vs. *hängen, hängte, gehängt* (Engl. *hang* transitive).[2]

These distinctions yield homonyms which to a large extent correspond to the ones traditionally distinguished, and the criteria used are operational and can be applied in most contexts.

For polysemy we use two criteria:

1. **government:** different government patterns (slot frames, subcategorization frames) yield different senses

   John passed the butter.
   John passed.

2. **consistency of semantic properties:** if conflicting semantic properties can be assigned to a lexical item, we must distinguish different senses, e.g.

   John passed the butter. (The object is moved)

[2] Note that in some cases variation on gender and inflection are possible without any semantic effect. Those cases should not lead to setting up different homonyms.

John passed the river. (The object is not moved)

While the first criterion can fairly easily be verified in sentential contexts (this is not to say there will not be problems in a number of cases), it is much harder to operationalize the consistency criterion. For two reasons:

1. It may be quite difficult to formulate the distinguishing semantic properties in a watertight way, and even if this can be achieved,

2. it may often be impossible to verify these properties in given contexts.

In conclusion it should be noted that homonyms and senses are defined within a language only, and not with recourse to some translation. This position differs from Lehrberger and Bourbeau (1988). Translations can be used as heuristics to guide the discrimination of senses, but they should never become part of the actual criteria used.

Syntactic relations and semantic compatibility

The government pattern of a lexeme can be described as a set of syntactic relations in which the lexeme functions as a head. To describe syntactic relations we use notions such as **subject, direct object, indirect object, prepositional object.** We do not use deep cases or θ-roles, as it is still unclear how these concepts can be properly defined (see Dowty, 1989).

That government patterns and semantic constraints on governed constituents are useful to distinguish senses and to determine translations has been clear to lexicographers for a long time, but they have failed to describe government patterns precisely let alone semantic constraints.

Seen from semantic interpretation, syntactic constructions can be characterized as

1. compositionally interpretable constructions,
2. constructions which can be interpreted using regular processes of metonymy,
3. lexicalized constructions,
4. ad-hoc metaphors (neither lexicalized nor interpretable by rule).

These different types of constructions must be considered when the semantic compatibility of their components is discussed, and the notions **compositionality, metonymy, lexicalization, and metaphor** must be defined. We will not discuss these definitions here, and just mention one for semantic compatibility.

Pairs of senses of lexemes which participate in a compositional syntactic relation are called semantically compatible with respect to that syntactic relation. We do not want to characterize semantic compatibility just in terms of sets of pairs of senses, but rather look for the most general description of modifiers of a given head, i.e. we look for the **semantic types** of modifiers.

Analytical methods

The classic linguistic methods which are available to us are: **substitution, deletion, permutation, augmentation, transformation, and translation.** The sources which are available are text corpora, dictionaries, informants, and introspection. All of these types of sources should be used to compensate the disadvantages of each type.

In order to determine semantic compatibility and translation relations we need to answer the following questions:

1. Does the lexeme a have more than one sense?

   This is confirmed if we can find hypernyms (supertypes) for a which if substituted for each other in some context yield either anomalous or contradictory expressions. Compare

   eine Spannung einstellen (to set a voltage)
   einen Menschen einstellen (to hire a person)

   It is not very difficult to find contexts where *Spannung* and *Mensch* and their supertypes cannot be substituted for each other and thus to establish two senses for *einstellen.*

2. Does the syntactic relation *r* hold between lexemes *a* and *b*?

3. Is *r(a,b)* compositional, metonymic, lexicalized, or metaphoric?

   If a construction is compositional, we expect ail of the tests mentioned to be applicable without yielding anomaly. Metonymic expressions may be just like compositional ones, except that certain rules such as **institution used for person** or any other one (see Fass, 1991) need to be applied. Lexicalized and metaphoric expressions usually show restrictions under the tests mentioned (for discussion see e.g. Weinreich, 1963).

4. Is *a'* a translation of *a* under the condition *r(a,b)* ?

   E.g., if *r(a,b)* is *acc(einstellen, Spannung)* then *a'* is *set* (and *b'* is *voltage)* according to available bilingual corpus data. This may not always be as clear cut: *richtige Spannung* could be *correct voltage* or *true suspense.*

In order to be able to represent semantic compatibility we need a lattice for word senses based on the hyponymy relation, and we need to record the most general modifiers in the government patterns of each word sense. When the hyponymy relation is defined in such a way that a sense *a* is a subtype of some sense *b* iff all potential compositional contexts of *b* are also potential contexts of *a* , then we are more in accordance with linguistic usage than traditional approaches (see Dahlgren and McDowell, 1989; Breidt, 1991).

Statistical methods

The idea of using statistics in MT dates back to the forties, but for obvious reasons - lack of appropriate computer power and of machine readable text corpora, etc. — it is just beginning to enter the field. Statistical analysis of natural language is faced with a number of problems of which we can address only a few here. Most notable is the issue of **representativeness** of text corpora (see Sebba, 1991, for a recent discussion). Kucera et al. (1967) and Erk (1972) noticed that word frequency strongly depends on domain and/or genre of texts. This dependency is substantially stronger for nouns than for verbs. Our own experiments confirmed these results. We used a 16 million token corpus of German newspaper texts coming from the domains economics, culture, politics, and general news, and counted the lemmatized verbs and nouns. A χ^2-test over the four domains yielded apart from significant differences in distribution that nouns are three times as domain sensitive as verbs, i.e. their mean χ^2 -value is three times as high as that for verbs (at a significance level of 1%).

Lemmatization (normalization of inflected word forms) often requires taking context into account[3] as does the recognition of homographs and senses. Parsing corpora syntactically would solve this to a large extent, but since parsing of large corpora is still problematic due to required grammatical coverage and processing time, simplified approaches have been looked for, e.g. use of tagged texts.

Language models based on Markov processes have proved quite successful in speech recognition, which encouraged Brown et al. (1990) to try the same approach for MT. They use a language model and a translation model which applies notions such as **fertility** and **distortion** of target words relative to source words. The parameters are computed on the basis of monolingual and bilingual corpora. The results achieved with this rather simple approach are

[3] For example, during lemmatization and counting of verbs we could not differentiate between *sein* as the auxiliary verb or as the possessive pronoun. Other examples were *sieben* as verb (to *sieve)* versus the number seven, or *äußerst* as the 2nd person form of *äußern (to utter)* versus the adverb *(utterly).*

not very impressive, so it is not too surprising that an enrichment of the model with further analytical assumptions is being attempted.

Since statistical methods are much more difficult to apply to semantic relations between words than to single words, Church and Hanks' (1990) work is important which used **mutual information** */(x,y)* of pairs of words *x* and *y* which occur in "windows" of varying sizes which are laid over a text, and show that this is an effective method to find semantic associations between words. The same measure can also be used to determine semantic compatibility as it was described above or to associate words with their translations in an aligned bilingual corpus, although in this area — as Church and Gale (1991) have shown — a different distance measure may be better suited.

Dagan et al. (1991) did similar statistical analyses using a text corpus of the target language and translation equivalents suggested in a bilingual dictionary. They first determined syntactic relationships by parsing sentences and then statistically evaluated the word associations. With this combined method they show in a preliminary experiment that an 18% improvement can be achieved over a method where the disambiguation information is based on the word frequency alone.

In our own approach we try to determine semantic constraints on translation relations by looking at monolingual and bilingual corpora. Monolingual data have been analyzed for verb noun cooccurrences within a sentence in order to determine semantic compatibility. Since we do not parse the sentences, we need to determine syntactic relationships underlying the cooccurrences where mutual information suggests a significant association. From the analytical model described above we are aware that lexical ambiguity and the other factors mentioned need to be accounted for. In a next step we plan to analyze semantic generalizations of word pairs found such that the semantic types of verb arguments can be empirically justified. Bilingual analyses will be based on these results.

We present a few examples having a high value for mutual information (*I* > 2.0), showing some problems for determining syntactic relations. *Zahl-ansteigen (number-increase, I* = 5.32, subject) is reliably found, even in sentences like

> Weder der Bierverbrauch *stieg* bisher spürbar *an* noch die *Zahl* der Betrunkenen. (Neither the consumption of beer *increased* noticeably up till now nor the *number* of drunk people.)

which are hard to parse.

German verbs with separable prefixes tend to distort the picture, as is shown by

Frage-werfen (throw-question, I = 2.20, semantically incompatible) vs.
Frage-aufwerfen (raise-question, I = 4.95, accusative object)

Words used as prefixes often also belong to a different part of speech, as in

[auf dem] Programm-stehen (be [on the] program, I = 3.32)
Programm-aufstehen (rise-program, I = 5.27, semantically incompatible)

Auf is a preposition governed by the verb *stehen* as well as the prefix *of aufstehen.* In

Aufgabe-ansehen (regard-task, I = 3.62, als-complement)

of the 14 occurrences found in our corpus only 8 showed a proper syntactic relation. In spite of these examples we found in most pairs having a high mutual information value that a plausible syntactic relationship is obvious.

With respect to domain sensitivity of verb-noun pairs we have found out that their average χ^2-value, even though still statistically significant, is less than a fourth of the value for verbs and less than a tenth of the value for nouns. This fact can be exemplified by

Rolle-spielen (play-role) χ^2 = 19.22 (critical value: 11.36) and highest I value in 3 out of 4 domains.

This means that investigations of this kind are less dependent on the nature of a given corpus than one might expect.

Conclusions

Translation relations can be constrained by a number of factors, one of them semantic compatibility of lexemes in given syntactic relations. In order to describe semantic compatibility the classical linguistic tests can be applied to text corpora and other sources. But statistical methods can be applied as well, which is demonstrated by the work of Dagan et al. (1991), Church and Hanks (1990), and by our own work. Dagan's and our work also show that enrichment of statistical models by further analytical assumptions can be very fruitful.

Analytical and statistical methods each have their strengths and weaknesses. Where successful, analytical methods provide the explanatory models all scientists look for. Where our analytical models do not suffice in their explanatory power or their level of abstraction, where the relevance of linguistic phenomena is hard to assess, and where the linguistic input is less than perfect, statistical methods are called for.

What is the proper division of labor between analytical and statistical methods is still not settled, but at any rate it should be clear that we need both. Maybe this can teach us to stay away from the extreme positions both of empiricism and rationalism and embrace the more moderate approach of pragmatism in linguistics.

Acknowledgement

We want to thank the speech recognition group (the SPRING project team) of the IBM Scientific Center Heidelberg for the permission to use their newspaper text corpora.

References

Bläser, B, A. Storrer, U. Schwall (1992): "A Reusable Lexical Database Tool for Machine Translation". Submitted to COLING'92.

Breidt, L. (1991): "Die Behandlung von mehrdeutigen Verben in der Maschinellen Übersetzung", *IWBS Report 158.* Stuttgart: IBM Deutschland.

Brown, P., J. Cocke, S. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, P. S. Roossin (1990): "A Statistical Approach to Machine Translation", *Computational Linguistics,* **16,** 79-85.

Chomsky, N. (1986): *Knowledge of Language.* New York: Praeger.

Church, K. W., P. Hanks (1990): "Word Association Norms, Mutual Information, and Lexicography", *Computational Linguistics,* **16,** 22-29.

Church, K. W., W. Gale (1991): "Concordances for Parallel Text", *Proc. 7th Annual Conference of the UW Centre for the New OED and Text Research.* Oxford, England.

Dagan, I., A. Itai, U. Schwall (1991): "Two Languages Are More Informative Than One", *Proc. ACL-91.*

Dahlgren, K., J. McDowell (1989): "Knowledge Representation for Common Sense Reasoning with Text", *Computational Linguistics,* **15,** 149-170.

Dowty, D. R. (1989): "On the Semantic Content of the Notion of 'Thematic Role'" , in: G. Chierchia, B. H. Partee, R. Turner (eds.) : *Properties, Types, and Meaning II.* Dordrecht: Kluwer. 69-129.

Fass, D (1991): "met*: A Method for Discriminating Metonomy and Metaphor by Computer", *Computational Linguistics,* 17, 49-90.

Erk, H. (1972): *Zur Lexik wissenschaftlicher Fachtexte, Verben - Bd 1, Substantive- Bd 2,* München: Hueber.

Kucera, H., W.N. Francis (1967): *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Leass, H. (1991): "Anaphora Resolution for Machine Translation: A Study", *IWBS Report 187.* Stuttgart: IBM Deutschland.

Lehmann, H. (1973): *Linguistische Modellbildung und Methodologie.* Tübingen: Niemeyer.

Lehrberger, J., L. Bourbeau (1988): *Machine Translation — Linguistic characteristics of MT systems and general methodology of evaluation.* Amsterdam: Benjamins.

McCord, M. (1980): "Slot Grammars", *Computational Linguistics,* 6, 31-43.

McCord, M. (1989): "Design of LMT: A Prolog-Based Machine Translation System", *Computational Linguistics,* **15,** 33-52.

McCord, M. (1991): "The Slot Grammar System", in: J. Wedekind and C. Rohrer, (eds.) : *Unification in Grammar,* MIT Press. (To appear).

Quine, W. V. O. (1953): "Two Dogmas of empiricism", in: W. V. O. Quine: *From a Logical Point of View.* New York: Harper and Row. 20-46.

Quine, W. V. O.: "Meaning and Translation", reprinted in J. A. Fodor, J. J. Katz (eds., 1963): *The Structure of Language.* Englewood Cliffs, NJ: Prentice Hall. 460-478.

Rimon, M., P. Martinez, M. McCord, U. Schwall (1991): "Advances in Machine Translation Research and Development in IBM", 3rd International MT Summit, Washington DC.

Sebba,M. (1991): "The Adequacy of Corpora in Machine Translation", *Applied Computer Translation,* Vol. 1, No. 1.

Weinreich, U. (1963): "On the Semantic Structure of Language", in: J. H. Greenberg (ed): *Universals of Language.* Cambridge, Mass.: MIT Press. 142-216.