From: AAAI Technical Report SS-93-02. Compilation copyright © 1993, AAAI (www.aaai.org). All rights reserved.

Combining Logic-Based and Corpus-Based Methods for Resolving Translation Mismatches

Megumi Kameyama
SRI International
AI Center and CSLI
333 Ravenswood Ave.,
Menlo Park, CA 94025
megumi@ai.sri.com

Stanley Peters
Stanford University,
Linguistics and CSLI
Stanford, CA 94305
peters@csli.stanford.edu

Hinrich Schütze
Stanford University
Linguistics and CSLI
Stanford, CA 94305
schuetze@csli.stanford.edu

Introduction

Very often in good quality translations, a target language expression (word, sentence or text) differs in meaning from the corresponding expression in the source language. Fortunately quality translation consists in approximating the meaning of a source document, for the same meaning often cannot be rendered exactly in the target language.

This impossibility is called *translation mismatch* by Kameyama, Ochitani, and Peters [4], where two of us discussed some factors involved in choosing an appropriate meaning approximation as required. The present paper continues that discussion, and draws conclusions regarding bi-lingual lexicons.

Languages differ in the concepts and real-world entities that they have words and grammatical constructs to describe. This is why translation must frequently be a matter of approximating the meaning of a source language text rather than finding an exact counterpart in the target language. A major cause of the inadequate quality presently achieved by machine translation (MT) is the inability to generate contextually appropriate approximations when no exact translation exists between the source and target languages. An MT system needs to recognize mismatches when it encounters them, and to resolve mismatches either by recovering implicit information from their context, as required by the target language, or by leaving out some information the source text gives.

English has lexical items like *picture*, *photograph*, *painting*, *water color*, *oil painting*, *drawing*, and *line drawing*, whose meanings are partially ordered by the relation of hyponymy (inclusion of meaning). In the same semantic field, Japanese has *e*, *syasin*, *suisaiga*, *aburae*, and *senbyou*. Only four words in each language have synonyms in the other language (i.e., words with the same meaning): *water color* (when it refers to the painting) and *suisaiga*, *oil painting* and *aburae*, *line drawing* and *senbyou*, *photograph* and *syasin*. The Japanese word *e* is more specific than English *picture*, as it does not apply to photographs; however, it is more general than *painting* and *drawing*, both of whose meanings it includes. To translate an English statement about pictures into Japanese, one needs to choose between *e* and *syasin*. To translate a Japanese statement using *e* into English, one often needs to choose between *painting* and *drawing*, though sometimes one must select *picture*.

A similar situation exists with verbs whose meaning (a property or relation) selects for certain types of argument. For example, *play* must be translated as different Japanese verbs depending on the kinds of object played — *hiku* (pull) for string instruments, *huku* (blow) for wind instruments, *tataku* (hit) for percussion instruments, *suru* (do) for sports, and *enziru* (perform) for theatrical pieces. When the object is *gakki* (musical instrument), the verb to be used is *hiku* (pull) for string instruments, so *hiku* is more general than *huku* or *tataku*. (In other words, *hiku* is the 'unmarked' choice of verb for playing musical instruments.) Here, the relative inclusion relationships among different types of playing are determined by the inclusion relationships among different types of objects played.

Translation mismatches exists even between closely related languages such as French and English. The word *drug* is more general in meaning than either *drogue* or *médicament*. The word *droit* is ambiguous, having one meaning related to *duty* and another related to *law* (the field of jurisprudence). The word *law*, also ambiguous, has a second meaning related to *loi*, which has no meaning in common with *droit*.

In this paper, we focus on the problem of lexical mismatches and propose a method for mismatch resolution that effectively combines logical and statistical analyses of word meanings. We then draw out the implications for lexicons to be used in MT.

Mismatch Resolution Using an Information Lattice

Kameyama, Ochitani, and Peters [4] proposed to analyze mismatches in an information lattice based on situation theory [2]. In this approach, pieces of information, whether they come from linguistic or non-linguistic sources, are represented as *infons* [3]. For an n-place relation P, $\langle\langle P, x_1, ..., x_n ; 1\rangle\rangle$ denotes the informational item, or infon, that $x_1, ..., x_n$ stand in the

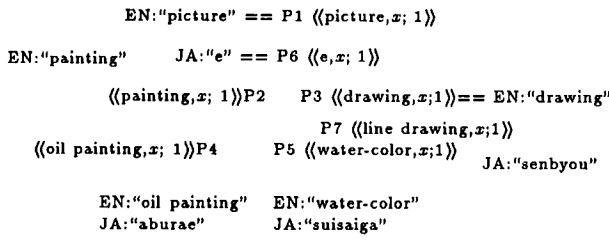

Figure 1: The "Picture" Sublattice

relation P, and $\langle\langle P, x_1, ..., x_n ;0\rangle\rangle$ denotes the infon that they do not stand in the relation. Given a situation s, and an infon σ, s $\models \sigma$ indicates that the infon σ is made factual by the situation s, read s *supports* σ .

Infons are assumed to form a distributive lattice with least element **0**, greatest element **1**, set I of infons, and "involves" relation $\Rightarrow$ satisfying:[1]

for infons σ and τ, if s $\models \sigma$ and $\sigma \Rightarrow \tau$ then s $\models \tau$

This distributive lattice $\langle I, \Rightarrow\rangle$, together with a nonempty set *Sit* of *situations* and a relation $\models$ on *Sit* $\times$ I constitute an *infon algebra* [1].

The information lattice represents logical aspects of the lexical and grammatical meanings expressed by source and target languages — inclusion relationships among related concepts and predicate-argument structures. The lexical meanings in this approach form an information sublattice, and a bilingual or multilingual lexicon to be used for translation will represent all the word senses of the given languages with all the known inclusion relationships made explicit.[2] Figure 1 (from [4]) is a lexical sublattice of some picture words discussed above, as an example of lexical mismatches between English (EN) and Japanese (JA). The picture words in question SIGNIFY (written ==) respective word senses (or properties) P1,...,P7 whose inclusion relations are captured with the INVOLVES relation ($\Rightarrow$).

High quality translations generally preserve exactly the same concept when possible, e.g., translate *water color* (P5) as *suisaiga* (P5). When the target language lacks a word or simple phrase for something in the source text, translation must either GENERALIZE the concept mentioned (e.g., translate *painting* (P2) as *e* (P6)) or SPECIALIZE it (e.g., translate *picture* (P1) as *e* (P6)). We have described these two general schemes for resolving mismatches using the notion of information flow [4].

This approach also applies to lexical mismatches where a verb's meaning depends on its argument types. In the *play* example discussed above, the English verb must be translated as different Japanese verbs depending on the kinds of object played. Recall that the relative inclusion relationships among different types of playing are determined by the inclusion relationships among different types of objects played. In general, for any n-ary relation P, the inclusion relationship of the arguments determine the inclusion relationship of infons of P — $\langle\langle P, x_1, ..., x_n ;1\rangle\rangle \Rightarrow \langle\langle P, y_1, ..., y_n ;1\rangle\rangle$ if $x_i \Rightarrow y_i$ for all $1 \leq i \leq n$.

Kameyama, Ochitani and Peters' 1991 proposal [4] was to use such a multilingual information lattice both to recognize mismatches and to resolve them. We can generalize their claim as follows: Given a linguistic form signifying a property P in the source sentence SL, the corresponding linguistic form in the target sentence TL should signify, in the descending order of preference,:

1. the same property P,
2. the closest property P′ such that P $\Rightarrow$ P′ (generalization),
3. the closest property P″ such that P″ $\Rightarrow$ P (specialization),
4. or some property Q (in the neighborhood of P) that is neither involved by P nor involves P.

When the linguistic form in question is a complex structure such as a sentence, the associated information is also a complex set of infons. Two sentences can convey the same complex information even if the pieces of information come from different kinds of linguistic sources. In other words, the same pieces of information can have DIVERGENT sources in the source and target sentences, and translation can still be said to preserve the same information. For instance, consider the translation from English *the red block is on the white block*[3] into Japanese *akai tumiki wa siroi tumiki no ue ni notte iru*. The spatial relation described with *on* loses the information of contact when it is described with *ue* ('on' or 'above'), but this lost information is regained by *ni notte iru* (be riding on), which specializes the relation *is on*. Thus the preferred strategy for resolving lexical mismatches is to allocate, if possible, the same pieces of information in different parts of the phonologico-morpho-syntactic structure of the sentence. In this process, some lexical mismatches are resolved with generalization, others are resolved with specialization, and yet others are resolved with some other strategy.

It is, however, not possible always to find an exact translation for a given sentence. In the above example of blocks, for instance, the singularity information

[1] We assume that the relation $\Rightarrow$ on infons is transitive, reflexive, and anti-symmetric after Barwise and Etchemendy.

[2] We set aside for now the question of how such a bilingual conceptual lattice would be constructed, except to say that it would draw on a combination of monolingual thesauri, bilingual dictionaries, and introspection of bilingual speakers. Monolingual thesauri outline monolingual conceptual lattices. Bilingual dictionaries outline the complex lexical mappings between the two languages. Bilingual speakers' introspection backs up the generalizations obtained from the first two sources. We would like to make this construction process semi-automatic, but how it can be done is still unclear.

[3] Assume that the context gives the information that the red block is on top of the white block.

of each block is totally lost in the natural translation in Japanese. We must, therefore, find a way to move within the information lattice to find an optimal translation.

Generalization is preferred because it is often a safer translation strategy than specialization — that is, it does not overcommit. For instance, translating a Japanese sentence *watasi wa kono e ga suki desu* (I like this 'e') into English *I like this picture* is safer than, say, *I like this painting unless* the discourse context specifies the referent of 'e' to be a kind of painting. A sound mismatch resolution strategy thus must take into account the information obtained from the discourse-specific context, so that the context allows specialization of mismatches without false overcommitments. In other words, the information associated with a sentence in discourse is not just its linguistic meaning out of context but is its linguistic meaning fleshed out in the given discourse context — what we call the UTTERANCE MEANING. In the next section, we will spell out (1) what is an utterance meaning and (2) what is closeness of information.

Approximating the Utterance Meaning

It is useful to distinguish between the linguistic meaning of an expression in a language and the utterance meaning expressed by using that expression on a particular occasion. An expression possesses its linguistic meaning independently of any context in which it may be produced or interpreted. The utterance meaning of any utterance is a function of the linguistic meaning of the expression uttered together with specific features of the context in which it is uttered. We have seen that a good translation of a word or sentence often has a different linguistic meaning than the original. Next we will show how the utterence meaning of source and target expressions in context are often much closer than their linguistic meanings.

The information carried by a sentence varies with context as a result of several causes.

1. The intended structure and meaning of an ambiguous sentence depends on the context in which the sentence is used.
2. The literal interpretation a given meaning yields is a function of indexical factors.
3. The focus, salience, and Gricean implicatures along with other communicative impacts determined by a given literal interpretation vary according to context.
4. The picture a sentence portrays in context adds to its communicative impact certain information obtained from logical implications, natural laws, and other contingently true generalizations about the domain of discourse.

The information given by a sentence in context is the result of the action of all these factors. This is the UTTERANCE MEANING of a sentence.

In Kameyama, Ochitani, and Peters (1991), we employed situated utterance representations (SURs) — four-tuples ⟨ DeT, PS, DiS, US ⟩ — to formalize sentences in context.

Described Situation Type (DeT) The way a certain piece of reality is, according to the utterance

Phrasal Situation Type (PS) The surface form of the utterance

Discourse Situation Type (DiS) The current state of the on- going discourse when the utterance is produced

Utterance Situation Type (US) The specific situation where the utterance is produced

The components of a SUR correspond to context and different aspects of its effects as follows. DeT corresponds to the output of item 2, PS to the structural part of the output of item 1, DiS to the contextual inputs of item 3, and US to the contextual inputs of items 1 and 2. The information given by a sentence in context, item 4, is not directly represented in a SUR.

When comparing a source sentence's SUR ⟨ DeT, PS, DiS, US ⟩ with the SUR ⟨ DeT′, PS′, DiS′, US′⟩ of its translation, the literal interpretations, DeT and DeT′, can be very different in the case of translation mismatches. Because translators take context into account, translation approximates between the context-dependent information (utterance meaning) that is the closure of DeT or DeT′ under Gricean and other communicative effects as well as under logical and natural laws and various contingent generalizations.

To show how similarity between such utterance meanings can be used to define the translation relation $\sim$ between source and target SURs, we make use of the notions defined in the next section.

Measuring Informational Distances

We propose to measure informational distances in terms of neighborhoods (open sets) in a topology of information:

DEFINITION. Item i_1 of information is as near to i_2 as i_3 is if every neighborhood (open set) containing both i_2 and i_3 contains i_1.

The appropriate characterization of open sets makes this definition equivalent to the condition that the information $i_2 \wedge i_3$ includes $i_1 \wedge i_2$.

The relevant notion of inclusion of information is given by the combination of logical implication ($\Rightarrow_l$), communicative principles ($\Rightarrow_c$), natural laws ($\Rightarrow_n$), and other contingently true generalizations about the domain of discourse ($\Rightarrow_d$). These heterogeneous principles combine to determine the utterance meaning — the information given by a sentence in context from the sentence's literal interpretation. We can define overall information inclusion $\Rightarrow$ as the union of all these different sources of informational inclusion:

DEFINITION. $\Rightarrow = (\Rightarrow_l \bigcup \Rightarrow_c \bigcup \Rightarrow_n \bigcup \Rightarrow_d)*$

Then the collection of open sets $\{i | i' \Rightarrow i\}$ for all items i' of information forms a topology, the appropriate one for approximating information in translation.

Direction and Distance of Mismatch

If the information of a source language sentence cannot be reproduced exactly with a sentence of the target language, each site of a mismatch between the two languages must be resolved to produce an overall adequate approximation. Mismatch resolutions can

1. lose context-independent information, becoming more general (generalization),
2. add context-independent information, becoming more specific (specialization), or
3. do neither, becoming incomparable (neither more general nor more specific).

Generalization chooses target information that includes the source information. Specialization chooses target information that is included in the source information. In our topology, 'generalizing' corresponds to choosing a target expression whose information lies in an open set containing the source information.

In translation mismatches, generalization is the more conservative strategy because it is guaranteed to preserve accuracy of information. However, the generalization should be slight, or else a good deal of information will be lost. We wish to explore the hypothesis that minimal generalization is always the best strategy in translation — that is, translators always aim to choose a target sentence that *in context* gives information which lies in a small open set around the information given by the source sentence in its context. We state the following principle:

PRINCIPLE. Minimal generalization is always the best strategy in translation.

The viability of this hypothesis depends crucially on recognizing it is context-dependent utterance meaning, not context-independent linguistic meaning, that the best translation generalizes. In the terms of the previous section, $\langle$ DeT, PS, DiS, US $\rangle \sim \langle$ DeT′, PS′, DiS′, US′$\rangle$ holds iff the literal interpretation DeT′ of the target utterance is in a sufficiently small open set around DeT.

The validity of this principle depends on two important facts about the context-dependent information. One is that the topology of information is context-dependent; the neighborhoods around an item of information shift in a regular way as $\Rightarrow_c$ and $\Rightarrow_d$ vary with context. We discuss this fact further in the next section.

The second crucial fact is that many apparent cases of mismatch resolution by specialization receive a regular explanation in terms of generalizing the context-dependent information given by the whole sentence. When a lexical item appears in an downward monotone position, for instance, in the scope of a negation, replacing it by a more specific item yields a sentence whose information generalizes that of the original. For example, consider the lexical mismatch around the 'clock' words in English and Japanese. Japanese *tokei* (time pieces) means either watches (worn time pieces) or clocks (not worn). Thus in translating *I have no clocks* into Japanese, it is not conservative to generalize 'clocks' into 'tokei', producing *watasi wa tokei wo motteinai* (I have no time pieces), since this specializes the information of the overall sentence and it is possible that the speaker has a watch. Depending on what open sets the context determines, one might instead specialize and translate *clock* into *mezamasidokei* (alarm clock), yielding the generalization *watasi wa mezamasidokei wo motteinai* (I have no alarm clocks) of the overall sentence's meaning.

Note that a good translation must be a *minimal* generalization in the given context. In the above 'clock' example, for instance, the context should place 'alarm clock' close to 'clock' since the source sentence does not specifically mention alarm clocks. Depending on the particular context, some other translation may be better. The next section discusses how context-dependent closeness of information depends on a given discourse.

Change of Topology with Context

One way to regard the reason why a word or sentence can have different best translations in different contexts is to consider as context-dependent the topology of literal information content, the topology varying especially with communicative effects ($\Rightarrow_c$) and with some domain facts ($\Rightarrow_d$).

While an expression can express different utterance meanings in different contexts, and an utterance meaning can be expressed by different expressions in different contexts, the topology of utterance meanings themselves is relatively fixed and independent of context. What we do here is rearrange in a context-dependent way the topology on literal interpretations — collapsing distinct ones that give rise to the same utterance meaning in a given context — so that the result is isomorphic to the topology of utterance meanings themselves.

Interestingly, only some generators of the topology vary with context; $\Rightarrow_l$ and $\Rightarrow_n$ are relatively constant. If we distinguish between atemporal and temporal determinants of the topology corresponding to the relative temporal stability, the most stable is logical implication ($\Rightarrow_l$) and the second most stable is the natural laws ($\Rightarrow_n$). We may consider them fixed in each translation instance. In contrast, communicative effects and perhaps some domain facts depend on things that change across translation instances.

Now each of these four types of information inclusion generates a topology by itself, and the topology we desire for approximating information is their product. So it is useful to factor it into components that are sta-

tionary and others that vary with time. It is, however, unclear exactly how to obtain these different kinds of topology, especially those that vary with the discourse progression that are supposed to contain information such as the subject matter, the speaker's purpose, attitudes and perspectives, and the currently salient entities. This is where we propose to make use of the statistical information obtained from alined bilingual corpora.

Mismatch Resolution using Sublexical Space

Another way to view the fact that good translations approximate the source's utterance meaning, and our theory of this, is as follows. With respect to lexical meaning, we can observe that each specialization of a word's linguistic meaning to an utterance meaning is associated with a particular type of utterance context. Suppose we could find a way to determine when the context of a given use gives rise to a particular utterance meaning; then we could choose a translation appropriate to that utterance meaning.

Schütze's work on word sense disambiguation is directly pertinent to this problem. His algorithm uses SUBLEXICAL REPRESENTATIONS [6] derived from a large text corpus. First, a collocation matrix is collected for several thousand high-frequency words. Each element $a_{i,j}$ in the matrix records how often words w_i and w_j co-occur in the corpus (in a sentence or in a window of, say, 100 words). A principal component analysis is then performed on the collocation matrix and on the order of one hundred principal components extracted. As a result, each word is represented by a vector of 100 components. The vector representations capture semantic and pragmatic regularities to the extent that collocations characterize the set of topics and contents in connection with which a word can be used.

The key information captured in the space is SEMANTIC SIMILARITY. The closest neighbors of a word in the space (according to some measure such as the cosine) are semantic cognates. The more distant two words are from each other in the space, the more different are the semantic fields they belong to. An approximate representation of the topical character of the context of a sentence can be computed by summing up the vectors of all words in the sentence (or computing their CENTROID). Disambiguation proceeds by clustering the vectors of all contexts that an ambiguous word occurs in, assigning senses to the clusters (by inspecting a few samples from the cluster) and disambiguating a new context of the ambiguous word depending on which cluster it is closest to. Disambiguation rates between 89% and 96% were achieved using this algorithm [6].

This framework can be extended to representing words from several languages in the same space with the help of an aligned corpus, for instance by collecting a matrix that records co-occurrence in the aligned corpus such that $a_{i,j}$ counts the number of sentence pairs with French word w_i occurring in the French sentence and English word w_j ocurring in the English sentence. (The bilingual sublexical space computed for this paper is based on a slightly different and computationally less expensive procedure.) The Canadian Hansard with Ken Church's alignment was used here.

Mismatch resolution can then be performed by searching the space for the word in the target language that is closest to the original word in the source language. If there is a one-to-one correspondence for a particular pair of words from source and target language, then they will occupy the same location in the space. In the case of a mismatch, the word in the source language doesn't have a direct translation, and it does not share its location with a word from the target language. Choosing the closest word as a translation, minimizes the mismatch between original and translation with respect to the semantic and pragmatic regularities that were automatically extracted by the principal component analysis.

A different way to view this proceeds by noting that the clusters of a given word's vectors in context space correspond to different TOPIC FIELDS in which the word is used — to contribute particular utterance meanings to a larger discourse. If the word's utterance meaning is unambiguously determined by its topic field at each occurrence, then determining the topic field should be valuable in figuring out which target word will have approximately the same utterance meaning in the given context. This observation will be important when we take up dictionaries for MT.

Lexical disambiguation and mismatch resolution

An ambiguous word can be disambiguated to different linguistic meanings as a function of context, just as a single linguistic meaning of a word can give rise to different utterance meanings in different contexts. Both phenomena (as well as their combination) cause contextual variation in what target expression is the best translation of a given source expression.

It is worth noting that to translate we needn't settle whether, for instance, the English word *drug* is ambiguous between 'medication' and 'substance that causes addiction,' or unambiguously means something that generalizes both these putative senses. Choosing the appropriate one of the two French words *drogue* (illegal drugs) and *médicament* (prescription drugs) as a translation can proceed similarly in terms of the topic fields obtained from corpora. The different utterance meanings of *drug* correlate with different topic fields, whether they arise from different linguistic meanings or the same one. Although it is sometimes difficult to sort out 'mere' differences of utterance meaning from genuine ambiguities of linguistic meaning, this need not impede translation on our approach.

If *drugs* is ambiguous between DRUGS1 (illegal drugs) and DRUGS2 (prescription drugs), sense disam-

biguation by topic spaces solves the problem of translating into French, as DRUGS1 translates into *drogue* and DRUGS2 translates into *médicament* in one-to-one fashion. If *drugs* has only one sense DRUGS, on the other hand, the topic spaces corresponding to its different utterance meanings helps choose the French translation. Either way, the two distinct topic fields for *drugs* are:

DRUGS1
(*addictheroinpusherssniffingthugsvialsnarcotic*)
DRUGS2
(*medicalPreventivephysicianspatientdiagnosesneurosurgery*)

It is possible, however, to identify true mismatches using a set of linguistic tests such as gapping and "how many" questions. For instance, we have determined that the clock-words in English and Japanese present a true mismatch. For a question *tokei wa ikutu motte imasu ka* (how many 'tokei' do you have?), the answer has to count all the watches and clocks (but not the stop watches). This shows that the Japanese word *tokei* does not have different senses for 'clocks' and 'watches'. Thus in translating the word *tokei* into English, we must perform a mismatch resolution rather than sense disambiguation in the context.[4]

Even though these statistical methods of corpus analysis do not identify the utterance meaning of an expression in context, they do provide a highly useful empirical constraint on the adequacy of detailed descriptions of the linguistic meaning of English and French words together with an account of how these give rise to utterance meanings in context. The utterance meanings predicted by such descriptions for English and French should be closely similar in precisely those cases where this analysis of corpora shows that English and French words or phrases are good translations of each other.

Synthesis

We have presented logic-based and corpus-based approaches to detecting and resolving mismatches in machine translation. They are needed because frequently no expression in the target language expresses all and only the information presented in the source language. Both the vector space and the information lattice define measures of relatedness helpful in selecting plausible candidates for the best compromise between the demands of specificity and correctness. However, both measures are problematic if used on their own.

In the lattice, the measure of relatedness is the number of rungs that have to be traversed to get from the meaning and structure of a source language expression to an expression in the target language. In the simplest case, adding or removing one infon corresponds to traversing one rung. Two important problems exist with this measure. First, some parts of the hierarchy may be analyzed more finely than others. A large number of arcs between two points in a dense part may be a comparable distance in semantic content to a difference of just one arc in a sparsely populated part of the lattice (cf. Resnik 1992 [5] who points out similar problems with WordNet). Second, world knowledge and the specifics of the text sort to be translated may play a crucial role in determining what translation is best. If a French novel speaks of a *gouache* in describing a room's decor, the best translation might well be a general term like *watercolor* if the technically correct word in the target language would be inappropriate for the general readership of a novel. In an art book, on the other hand, the technical term would be required. But the linguistic knowledge that the lattice is based on does not offer any clues as to how the measure of semantic relatedness could be adapted to different text sorts.

In the space, the measure of relatedness is the cosine of the angle between two vectors representing contexts or words. The vectors are based on the collocational patterns of words where context words a distance of 5 away from the word being classified are as important as context words at a distance of 30. Furthermore, syntactic structure is ignored. As a result, the vectors represent well what the word is used to talk about; but they contain very little information about what is actually said. For example, the closest neighbor of French *vert* is *red*, not *green*. Antonyms often end up as close neighbors. Oversimplifying only slightly, one can say that the propositions p and $\neg p$ occupy the same location in the space.

The crucial observation is that the two representational schemes get at INDEPENDENT aspects of topic-content. Whereas the space is finely tuned to the pragmatic factors affecting the composition of text, the lattice represents predicate-argument structure and logical relationships that are necessary for a description of what the proposition expressed is (vs. what the entities and situations talked about are). So combining the two measures of relatedness should give us the best of both worlds: logical clarity as well as pragmatic sensitivity to the particulars of the application text.

Combining these schemes seems natural for two reasons. Encoding a dictionary by hand is feasible with the investment of a couple of person-years. Representing all world knowledge necessary for a given task is much more labor-intensive. Secondly, different world knowledge is needed for each new application whereas the linguistic properties of texts are more stable across different genres. Using a trainable, albeit simplistic, component for world knowledge and an expensive, but theoretically well-founded component, for linguistic knowledge could thus prove to be the optimal combination.

Implications for MT lexicons

Linguists usually regard the semantic task of a lexicon to be stating the linguistic meanings of the words of a language, and the job of semantic rules to be stating how the linguistic meaning of syntactically generated

[4] We do not know yet what topic spaces *tokei* would show.

phrases and sentences are composed from the linguistic meanings of the lexical items they include. Lexicographers have taken varying stands on what meanings should be entered in a dictionary. Roughly speaking, the so-called splitters favor listing each possible utterance meaning of a word separately in the dictionary (while indicating similarities among related utterance meanings, of course). So-called lumpers, on the other hand, would list in one entry a single linguistic meaning that gives rise in context to multiple utterance meanings.

Bilingual dictionaries list source words with (in principle) all the various target language words they translate to. One problem in using these dictionaries is that they often give inadequate information to determine which translation is best in a given context.

Our techniques permit the construction of a bilingual dictionary in which each word is entered by associating each topic field in which it is found with the word (or words) of the other language that translate it in that context. The topic fields can be constructed automatically from a large corpus, and the topic-specific translations can also be constructed automatically from a corpus of translations. Such an automatically constructed bilingual dictionary for MT can be seen as a natural generalization of traditional bilingual dictionaries, which attempt to give crude topic fields (e.g., 'military', 'law', etc.) as a basis for choosing the translation of a word.

It may be that words cluster into more regions of context space than they have utterance meanings. This need not prevent the compilation of useful bilingual dictionaries, however. By employing a characterization of contexts grounded in an appropriate corpus of a language, and a large enough corpus of translations to find translation pairs in a wide range of topic fields, we may reasonably hope to find the occurrences needed for compilation of a dictionary that generalizes adequately to new documents that are sufficiently like the corpus.

While such a bilingual lexicon will not meet linguists' goals of representing each word's linguistic meaning in a unified way, it will nevertheless provide crucial empirical data for testing the correctness of linguistic lexicons when they become capable of predicting utterance meanings in terms of linguistic meanings and contextual factors. Whatever the contextual facts may be in a given topic field, the linguistic lexicon should predict an utterance meaning that is close to the one it predicts the best translation word has in the same context.

References

[1] Barwise, Jon and John Etchemendy. 1990. Information, Infons, and Inference. In Cooper, Robin, Kuniaki Mukai, and John Perry (eds) 1990. *Situation Theory and Its Applications*, Volume 1, CSLI Lecture Notes Number 22. Stanford: CSLI Publications, 33-78.

[2] Barwise, Jon, and John Perry. 1983. *Situations and Attitudes*. Cambridge, MA: MIT Press.

[3] Devlin, Keith. 1991. *Logic and Information*. Cambridge, England: Cambridge University Press.

[4] Kameyama, Megumi, Ryo Ochitani, and Stanley Peters. 1991. Resolving translation mismatches with information flow. In *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics*, Association of Computational Linguistics, 193-200.

[5] Resnik, Philip. Forthcoming. WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery. In Weir, Carl; Abney, Stephen; Grishman, Ralph; Weischedel, Ralph (Eds.). Forthcoming. *Statistically-Based NLP Techniques*. AAAI Press.

[6] Schütze, Hinrich. 1992. Dimensions of meaning. In *Proceedings of Supercomputing*.