AMTA-96 Tutorial

THE LEXICON FOR MT

Ann Copestake

aac@csli.stanford.edu

CSLI, Stanford University

Overview

What do MT lexicons look like?

- Some lexicalist approaches to MT
- What information is in the lexicon?

How can lexicons be acquired?

- Machine readable dictionaries
- Corpus-based approaches

A short case study

- Multilingual Web authoring tool

Aim: to describe some recent research and its possible broader application (NOT to give a systematic overview of all approaches or to discuss existing commercial systems).

PART I: What do MT lexicons look like?

Minimal MT lexicon: pairing of words (and phrases) in source and target languages. Translation is lexical lookup.

Elaborations: add information (statistical, morphological, syntactic, semantic, discourse, world knowledge) and complicate processing to:

1. Address the word ambiguity problem
2. Produce 'good' output, for example:
   - word ordering, agreement etc,
   - addition of information implicit in source language
   - avoid unintended readings
   - use natural collocations etc

In many MT approaches, most of the information in a broad-coverage system is lexical.

Design principles for hand-coded lexicons

Minimize content in order to minimize acquisition problem.

Declarativity and human readability with compilation into a machine-friendly representation.

Modularity so components are reusable: e.g. distinct monolingual and transfer lexicons.

Capture generalizations with inheritance (and lexical rules etc). Avoids errors, easier to maintain and expand.

Underspecification to reduce disambiguation: e.g. *Zeitung* and *newspaper*.

Some recent lexicalist approaches

Lexicalist semantic transfer

- Transfer rules operate on semantic structures which are close to lexical structure: i.e. predicates close to word names, transfer on underspecified representations
- e.g. SRI BCI/SLT system: Alshawi et al, 1991; Verbmobil: Dorna and Emele, 1996
- Most (potentially all) transfer rules are equivalences between predicates corresponding to individual words

Lexical sign transfer

- Transfer rules operate on bags of instantiated lexical signs
- main development at Sharp Research, UK: Whitelock, 1992; Poznanski et al, 1995: also e.g. Trujillo, 1995; Popowich, 1994
- Bilexicon relates one or more TL lexical signs to one or more SL lexical signs

Lexicalist semantic transfer

1. Conventional parsing yields (underspecified) logical form (e.g. QLF)
2. Transfer rules relate parts of SL LF to a TL LF
3. Generation proceeds from TL LF

Kim likes Sandy $\leftrightarrow$ Sandy gusta a Kim

```
[pres,
  [like,
    q_term(<t=quant, n=sing>,A,[event,A]),
    a_term(<t=ref,p=name>,
              B,[ name_of, B, kim ]),
    a_term(<t=ref,p=name>,
              C,[name_of, C, sandy])]]

trans([like1,tr(ev),tr(subj),tr(obj)] ==
        [gustar1,tr(ev),tr(obj),tr(subj)]
```


Spoken language translator (SLT) system

- Used by SRI in domain-specific speech-to-speech translation system between several language pairs
- Can reuse grammar, lexicon, parser and generator developed for other applications
- QLF underspecifies quantifier scope, doesn't resolve pronoun reference
- Heavy use of *packed* representations to avoid combinatorial explosion in parses and transferred LFs
- Uses statistical information for ambiguity resolution (*hybrid transfer*, see later)
- Transfer rules from L1 to L2 and L2 to L3 can be composed to give useful initial set of L1 to L3 rules (see later)
- Relatively low effort to change domain / add new language pairs

Lexical sign transfer

Logical form equivalence problem: how do you guarantee that your grammar will generate from the output of transfer?

'Mismatch' at LF
Suppose LF for *fierce black cat* is:
$\lambda x[\text{fierce}(x) \wedge (\text{black}(x) \wedge \text{cat}(x))]$
straightforward transfer to Spanish:
$\lambda x[\text{feroz}(x) \wedge (\text{negro}(x) \wedge \text{gato}(x))]$ but Spanish grammar for *gato negro y feroz* produces:
$\lambda x[\text{gato}(x) \wedge (\text{negro}(x) \wedge \text{feroz}(x))]$

Phrasal transfer
e.g. transfer rule for English:*young bull* $\approx$ Spanish:*novillo* should apply to *young black bull.*

Whitelock et al's solution: transfer operates on instantiated lexical signs.

Schematic structure for *Das ist schlecht*

$$\begin{bmatrix}
\text{phon} & \left\langle \textit{das ist schlecht} \right\rangle \\
\text{sem} & \boxed{1}\begin{bmatrix} \text{pred} & \text{and} \\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{deixis}\\ \text{arg} & \boxed{2}\end{bmatrix} \\ \text{arg2} & \boxed{4}\begin{bmatrix}\text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{schlecht_state}\\ \text{event} & \boxed{3} \\ \text{arg1} & \boxed{2}\end{bmatrix} \\ \text{arg2} & \boxed{5}\begin{bmatrix}\text{pred}&\text{present}\\ \text{arg1} & \boxed{3}\end{bmatrix}\end{bmatrix}\end{bmatrix} \\
\text{dtrs} & \begin{bmatrix} \ldots \begin{bmatrix}\text{phon} & \left\langle \textit{das}\right\rangle \\ \text{sem} & \boxed{1}\end{bmatrix} \\ \ldots \begin{bmatrix}\text{phon} & \left\langle \textit{schlecht}\right\rangle \\ \text{sem} & \boxed{4}\end{bmatrix} \\ \ldots \begin{bmatrix}\text{phon} & \left\langle \textit{ist}\right\rangle \\ \text{sem} & \boxed{5}\end{bmatrix}\end{bmatrix}
\end{bmatrix}$$

Source bag

$$\begin{bmatrix}\text{phon} & \left\langle \textit{ist}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{present}\\ \text{arg1} & \boxed{3}\end{bmatrix}\end{bmatrix}
\quad
\begin{bmatrix}\text{phon} & \left\langle \textit{das}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{deixis}\\ \text{arg} & \boxed{2}\end{bmatrix}\\ \text{arg2} & \end{bmatrix}\end{bmatrix}$$

$$\begin{bmatrix}\text{phon} & \left\langle \textit{bad}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{schlecht\_s}\\ \text{event} & \boxed{3}\\ \text{arg1} & \boxed{2}\end{bmatrix}\\ \text{arg2} & \end{bmatrix}\end{bmatrix}$$

(Note: schlecht_state abbreviated to schlecht_s)

Bilexical entries

$$\begin{bmatrix}\text{phon} & \left\langle \textit{das}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{deixis}\\ \text{arg} & \boxed{2}\end{bmatrix}\\ \text{arg2} & \end{bmatrix}\end{bmatrix}
\leftrightarrow
\begin{bmatrix}\text{phon} & \left\langle \textit{that}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{deixis}\\ \text{arg} & \boxed{2}\end{bmatrix}\\ \text{arg2} & \end{bmatrix}\end{bmatrix}$$

$$\begin{bmatrix}\text{phon} & \left\langle \textit{schlecht}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{schlecht_s}\\ \text{event} & \boxed{3}\\ \text{arg1} & \boxed{2}\end{bmatrix}\\ \text{arg2} & \end{bmatrix}\end{bmatrix}
\leftrightarrow
\begin{bmatrix}\text{phon} & \left\langle \textit{bad}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{bad_s}\\ \text{event} & \boxed{3}\\ \text{arg1} & \boxed{2}\end{bmatrix}\\ \text{arg2} & \end{bmatrix}\end{bmatrix}$$

$$\begin{bmatrix}\text{phon} & \left\langle \textit{ist}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{present}\\ \text{arg1} & \boxed{3}\end{bmatrix}\end{bmatrix}
\leftrightarrow
\begin{bmatrix}\text{phon} & \left\langle \textit{is}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{present}\\ \text{arg1} & \boxed{3}\end{bmatrix}\end{bmatrix}$$

Target bag

$$\begin{bmatrix}\text{phon} & \left\langle \textit{is}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{present}\\ \text{arg1} & \boxed{3}\end{bmatrix}\end{bmatrix}$$

$$\begin{bmatrix}\text{phon} & \left\langle \textit{that}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{deixis}\\ \text{arg} & \boxed{2}\end{bmatrix}\\ \text{arg2} & \end{bmatrix}\end{bmatrix}$$

$$\begin{bmatrix}\text{phon} & \left\langle \textit{bad}\right\rangle \\ \text{sem} & \begin{bmatrix}\text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{bad\_s}\\ \text{event} & \boxed{3}\\ \text{arg1} & \boxed{2}\end{bmatrix}\\ \text{arg2} & \end{bmatrix}\end{bmatrix}$$

Schematic structure for *That is bad*

$$\begin{bmatrix} \text{phon} & \left\langle \textit{that is bad}\right\rangle \\ \text{sem} & \boxed{6}\begin{bmatrix} \text{pred} & \text{and} \\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{deixis}\\ \text{arg} & \boxed{2}\end{bmatrix} \\ \text{arg2} & \boxed{7}\begin{bmatrix} \text{pred} & \text{and}\\ \text{arg1} & \begin{bmatrix}\text{pred} & \text{bad\_s}\\ \text{event} & \boxed{3}\\ \text{arg1} & \boxed{2}\end{bmatrix}\\ \text{arg2} & \boxed{8}\begin{bmatrix}\text{pred}&\text{present}\\ \text{arg1} & \boxed{3}\end{bmatrix}\end{bmatrix}\end{bmatrix} \\ \text{dtrs} & \begin{bmatrix} \ldots & \begin{bmatrix}\text{phon} & \left\langle\textit{that}\right\rangle\\ \text{sem} & \boxed{6}\end{bmatrix} \\ \ldots & \begin{bmatrix}\text{phon} & \left\langle\textit{bad}\right\rangle\\ \text{sem}&\boxed{7}\end{bmatrix}\\ \ldots & \begin{bmatrix}\text{phon}&\left\langle\textit{is}\right\rangle\\ \text{sem}&\boxed{8}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

Lexical sign transfer

- Assumes absolute lexicalism (i.e. all semantic information must have source in string)
- Main advantages: bilingual lexicons just involve relating monolingual signs, avoids (some aspects of) logical form equivalence problem
- Potential disadvantages: less modularity than semantic transfer, absolute lexicalism, efficient techniques involve restrictions on monolingual grammars
- Possible long term disadvantage: no possibility of doing reasoning to construct a revised TL LF

Flat semantics (Verbmobil)

Observation: LF equivalence problem can be alleviated by choice of the syntax of semantic representation:
E.g. instead of *fierce black cat* being:
$[\text{fierce}(x) \wedge (\text{black}(x) \wedge \text{cat}(x))]$ an equivalent representation is:
$_\wedge[\text{fierce}(x), \text{black}(x), \text{cat}(x)]$

- Need *label*s (UDRS) or *handel*s (MRS) for partial scope
- Bag of semantic structures (rather than bag of signs)
- Efficient techniques for transfer described in Dorna and Emele, 1996
- Generation could use Shake'n Bake algorithms, chart generation (Kay, 1996) or more conventional generators
- Verbmobil transfer uses conditions on transfer rules rather than statistics

Lexicalist approaches: concluding points

Main point: formally adequate approach to MT is possible with lexicalist transfer. So the transfer lexicon looks like a bilingual wordlist with addition of mapping between arguments, plus symbolic or statistical conditions on applicability.

- Statistical MT with transfer operating on dependency structures (Alshawi, 1994, 1996) is also lexicalist
- Lexicalist approaches are capable of dealing with complex head-switching examples (such as *swim across the river / cruzó el río nadando*) and capturing generalizations about them (see later)

Simplified flat semantic transfer equivalences used in the next section, but examples can be reformulated in other approaches.

Contents of MT lexicons

Morphology and syntax in monolingual MT lexicons are usually similar to general NLP.

Two main issues specific to MT lexicons in a transfer approach:

1. How can lexical semantics be exploited to make statement of transfer rules easier?
2. What information should the bilexicon contain, apart from simple word/sense/sign relationships?

Partial interlingua in semantic transfer approaches: tense and aspect

Rather label tense by PAST etc, interlingual approach analyses the meaning. For example: distinguish speech time (S), event time (E), reference time (R) (Reichenbach, 1947)

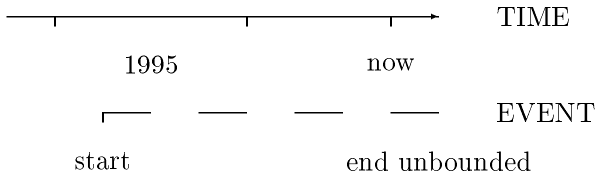
Tomorrow Kim will have got the email

S = now, R = tomorrow, E < R

Point-like vs extended events:

Since 1995 I have worked on Archimedes
Seit 1995 arbeite ich an Archimedes

Tense mismatch, but underlying situation is the same:



Lexical generalizations

Kim swam across the river
Kim cruzó el río nadando

Semantic forms are not equivalent:

$$\text{swim}(e1, K), \text{across}(e1, x), \text{river}(x)$$

$\text{cruzar}(e2, K, x), \text{rio}(x), \text{WHILE}(e1, e2), \text{nadar}(e1, K)$

(WHILE is for semantics of gerundive, but interpretation not relevant). Transfer rule required is:
$\text{swim}(e, w), \text{across}(e, z) \leftrightarrow$
$\text{nadar}(e, w), \text{cruzar}(e', w, z), \text{WHILE}(e, e')$

Generalize: assume PATH, MOVE-PATH, MOVE are (non-language specific) generalizations of the classes of predicate that include across, cruzar and swim, respectively.
$\text{swim}(e, w), \text{PATH}(e, z) \leftrightarrow$
$\text{nadar}(e, w), \text{MOVE-PATH}(e', w, z), \text{WHILE}(e, e')$

$$\text{MOVE}(e, w), \text{across}(e, z) \leftrightarrow \text{cruzar}(e', w, z)$$

Metalevel patterns

$$\text{MOVE}(e, w) \leftrightarrow \text{MOVE}(e, w)$$
$$\Downarrow$$
$\text{MOVE}(e, w), \text{PATH}(e, z) \leftrightarrow$
$\text{MOVE}(e, w), \text{MOVE-PATH}(e', w, z), \text{WHILE}(e, e')$

Thus, given $\text{swim}(e, w) \leftrightarrow \text{nadar}(e, w)$ we would also automatically get:
$\text{swim}(e, w), \text{PATH}(e, z) \leftrightarrow$
$\text{nadar}(e, w), \text{MOVE-PATH}(e', w, z), \text{WHILE}(e, e')$

Could also use an interlingual notion of spatial prepositions and path components:
e.g. semantics of *across* is $\text{CROSS-DIR}(e, z)$, semantics of *cruzar* is
$\text{MOVE}(e, w) \wedge \text{CROSS-DIR}(e, z)$ and no separate transfer rule is necessary. Plausible, because apparently limited range of spatial prepositions and types of path.

Conditional statements in the transfer lexicon

“Easy” examples:

- *eat* $\leftrightarrow$ *essen* / *fressen* depends on whether subject is human or other animal (except can be used of human to imply animal-like eating)
- *teacher* $\leftrightarrow$ *maestro* / *maestra* depends on gender (might be extremely difficult to resolve)

Difficult examples:

- *kommen* $\leftrightarrow$ *come* / *get to* etc depends on complex factors about location and ‘ownership’ of space
- *Termin* $\leftrightarrow$ *meeting* / *appointment* depends on factors like relative social status of participants

What should be encoded in the transfer lexicon? What should we use statistics for? What should the monolingual grammar do?

Hybrid transfer in SLT: 1

Even in limited domains, words often have multiple translations: for example, (at least) five translations of *on* into French in ATIS:

fly to Boston on Delta
aller à Boston avec Delta

information on ground transport
des renseignements sur les transports publics

a meal on that flight
un repas à bord de ce vol

the aircraft which is used on this flight
l’avion qu’on utilise pour ce vol

leave on Monday
partir le lundi

Hybrid transfer in SLT: 2

- transfer preferences are domain specific - *sur* is not the most frequent translation for *on* in the ATIS corpus.
- conditions are difficult to encode
- use statistical preferences to make choice

Unification-based transfer rules

```
trule([eng,fre], on1 >= sur1).
trule([eng,fre], on1 >= avec1).
```

Statistical preferences based on semantic triples

Positive triples:

```
(aller avec <airline>)
(renseignement sur transport)
```

Negative triples:

```
(aller sur <airline>)
(renseignement avec transport)
```


Word choice in monolingual grammar

We will meet **on** Thursday
We will meet **in** the morning
We will meet **on** Thursday morning

Choice of preposition ideally should be dealt with in monolingual grammar, because

* We will meet in Thursday

(bad sentence in any context where Thursday refers to a day of the week).

Interlingual notion of temporal location prepositions: English generator given e.g. temp-loc(e, x), day-of-week$(x, \text{Thursday})$, generates *on Thursday* because of restrictions in monolingual lexical entries.

Limited amount of work on encoding such restrictions in most linguistic frameworks.

Summary

- Lexical semantics can simplify statement of transfer equivalences
- Possible to use partial interlingua in a predominantly transfer based approach
- Relate patterns across languages without assuming meaning is identical
- Statistical vs symbolic techniques for encoding conditions on transfer
- Some word choice restrictions should be stated monolingually

But practical considerations:

- No point encoding complex conditions which cannot be resolved automatically
- No point assuming lexical information if it will be used very infrequently or can't be effectively acquired

PART II: Constructing lexicons for MT

1. Hand coding
2. Reuse of existing resources
3. Machine readable dictionaries
4. Corpus-based approaches

- acquisition of monolingual and bilingual lexicons involves different techniques but linking them is difficult if words have multiple senses
- can mine resources for particular type of information, rather than attempting to construct entire lexicon

Reusability

Primary problem is building **usable** lexicons: if it's usable then it's reusable!
But:

1. Procedural encoding (bad engineering?)
2. Lack of documentation, especially for semantics
3. Cost/benefit - not worth reusing a 100 word lexicon
4. Domain specificity
5. Legal issues

Many existing resources: wordlists (e.g. software companies), COMLEX, MRDs, Wordnet, EDR ...

Reusability by composing transfer rules

Translating between N languages requires $N(N-1)$ sets of transfer rules.
Can we compose transfer rules from L1 $\rightarrow$ L2 with those from L2 $\rightarrow$ L3 to get L1 $\rightarrow$ L3?

Rayner et al, 1996

- SLT system, ATIS domain
- Swedish $\rightarrow$ English with English $\rightarrow$ French
- Usable initial set (82% acceptable translations on test set)
- Cuts development time substantially
- Problems: automatically generated L2 differs from natural L2 (even if 'grammatical'); overgeneration (e.g. compounding rules); preference problems when combining statistical scores.

Machine readable dictionaries

1. Long history of research (see bibliography)
2. Limited explicit syntactic information, except in English learners' dictionaries (OALD, LDOCE, Cobuild, CIDE)
3. Noun definitions can be analysed to derive taxonomies: inheritance hierarchies for semantic information.
4. Less work on verbs (but e.g. Sanfilippo and Poznanski, 1992; Dorr et al,1995)
5. Little published work on bilingual lexicons (Copestake et al, 1995; Klavans and Tzoukermann, 1996)

Deriving monolingual semantic information

Categorize lexical entries using predefined linguistically motivated classes.
For nouns, the main source of information is automatically derived taxonomies.

Sauternes a type of sweet gold-coloured French wine

1. find *genus term* (*wine*)
2. disambiguate genus term wrt MRD senses
3. find the new entry's genus term (*drink*^{2} 1) etc, etc
4. associate taxonomy nodes with linguistic classes (encoded as types) — e.g. **comestible_substance**
5. information manually associated with type is inherited by lexical entries
6. refine classification using *differentia*

ACQUILEX experiment on building multilingual LKB

- Monolingual lexical fragments were constructed semi-automatically for four languages (English, Spanish, Dutch and Italian) from MRDs
- Derive *translation links* to form a multilingual lexical knowledge base (LKB)
- Link lexical entries corresponding to word senses in Spanish and Dutch to the English entries (which were more detailed)
- Use bilingual dictionaries to give candidate sets of lexical entry pairings, disambiguate using information from taxonomies
- Main problems: coverage of bilinguals not matching existing lexicons, phrasal translations

Example of sense correspondence problem for Dutch *doos* (Vossen, 1993)

VDmono sense	VDE sense	VDE translation	LDOCE sense
01⟨*object*⟩	01⟨*object*⟩	box	1 *container*
			2 *the contents*
			3 *room*
			4 *protector*
			5 *television*
02⟨*toilet*⟩	02⟨*toilet*⟩	loo (BE)	*lavatory*
(infrml)	(infrml)	lav (BE)	GAP
		john (AE)	01 *man*
			02 *lavatory*
		can (AE)	...
GAP	03⟨*jail*⟩	clink (BE)	...
	(infrml)	junk (BE)	...
		the can (AE)	...

Words in italics are to indicate the senses involved (angle brackets indicates English gloss)

Evaluation of ACQUILEX experiment

1. For subsets tested, almost no 'bad' links
2. Partial translation links generated if translation was a phrase (link to head of the phrase), or if no translation for word (link to translation of the closest ancestor in taxonomy)
3. % linked vary from 60% (English food subset) to 95% (Spanish drink subset) but partial links used in high percentage of cases of Spanish
4. Low % using bilingual dictionary 'in reverse' because bilinguals tend to only give more frequent forms as translations (e.g. *cider* rather than *cyder*)

Phrases used as translations in bilingual dictionaries

Often culturally specific:

- *Penedés*, *Rioja*, *Malaga* (local wines)
- *dubbelgebeide*, *bessejenever*, *moutjenever* (types of gin)

Phrases are not necessarily appropriate as a translation:

- He went into the bar and ordered a gin made with twice the normal quantity of juniper berries.
- He went into the bar and ordered a gin.
- He went into the bar and ordered a *dubbelgebeide*.

Hard-wiring a phrase as a translation is probably not the best option.

Evaluation of MRDs for MT

- ACQUILEX work only derived fragments, so evaluation in real MT system is not possible
- Best possible case for Verbmobil dialogue investigated manually (very small scale experiment)
- Deutsch-Englisch Langenscheidts Taschenwörterbücher (65000 words/phrases, relatively good coverage for conversations)
- Ignored closed class words and non-content words like *denn* (which are often left untranslated)
- About 25% of remaining terms were either not in (all were phrases or compounds for which a compositional translation did not work) or did not have a good translation for that use: e.g. *Termin* is given translations *date* and *appointment* etc, but not *meeting* which is often the appropriate translation in Verbmobil

MRDs: pros and cons

Advantages:

- dictionaries are used for manual lexicon construction
- broad coverage
- a labour-saving resource to augment a manually constructed core lexicon

Disadvantages:

- time-consuming to process initially
- highly variable quality
- different dictionaries require different strategies
- bilinguals inadequate even for **humans** (ambiguities, little frequency information, obsolete or offensive terms, poor coverage of idioms)
- publishers' IPR

Corpus-based acquisition techniques

- Test corpus essential for hand-coded lexicons
- Semi-automatic acquisition techniques: range from simple use of concordances to post-filtering of automatically acquired material
- Automatic acquisition of monolingual information: taggers which cope with unknown words, subcategorisation frames (e.g. Brent, 1991, 1993; Manning, 1993, Ushioda *et al.*1993, Briscoe and Carroll, 1995) semantic information (e.g. Ribas, 1994) etc
- Bilexicon acquisition from aligned corpora

Automatic acquisition of bilingual lexicons: the basic technique

1. Find some existing translated text (most experiments done on Canadian Hansards)
2. Align texts (by sentence or other unit)
3. For each word W1 in source language find set of sentences S1 in which it occurs and look at the frequency with which word W2 in target language occurs in sentences S2 which have been aligned with S1
4. For languages such as Chinese, an additional word segmentation step may be required, but see Wu and Xia, 1995

The alignment problem

Problems: sentences don't match 1:1, parts of text aren't translated, identifying sentences may be difficult/impossible.

- Alignment based on lexical correspondences extracted from bilingual dictionaries: Kay and Röscheisen, 1988; Catizone et al, 1989.
- Alignment of sentences based on their lengths (in characters, Gale and Church, 1991; or in words, Brown et al, 1991). Empirically better than lexical method, though Wu and Xia (1995) get improved results by using some lexical cues
- *char_align* method (Church, 1993) looks for character sequences which are the same in source and target
- Fung, 1995 — noun translations without sentence alignment (but coverage/precision not good)

Filtering

- Usual method is to construct lexicon based on co-occurrences of L1 and L2 words in parallel regions of text, assuming independence between word pairs – but errors occur due to indirect associations
- "Lexicon" constructed for statistical MT (e.g. Brown et al, 1990) not directly suitable for symbolic use: very long tail of low probability translations (mostly wrong)
- Most reported high precisions in automatic construction of bilingual lexicons are achieved by 'cherry-picking' top fraction of derived lexicon
- Melamed (1996: this conference) demonstrates high precision and high coverage by iterative filtering (but limited so far to 1:1 translations)

Collocations

- **collocation**: two or more words co-occur with a greater-than-expected frequency: e.g. *collective agreement*, *gays and lesbians*, *take . . . steps* — examples from Smadja et al, 1996
- May be non-compositional in meaning and/or have preferred phrasing (e.g. not usual to say *lesbians and gays*)
- Smadja (e.g. 1993) developed XTRACT: a tool for finding monolingual collocations
- Smadja et al (1996) take collocations identified by XTRACT and acquire translations from sentence aligned corpus, by iteratively identifying word sequences correlated with source and determining globally best: e.g. *official languages* gives best candidates *officielles*, *officielles langues*, *honneur officielles langues* etc

Corpora: pros and cons

Advantages:

- essential tool for human lexicographer
- domain-specific terminology and translations
- frequency information
- possibility of addressing ambiguity problem

Disadvantages:

- large scale parallel corpus may not exist for particular domain / language pair
- automatic extraction unproven in full systems: current methodology maybe too geared to producing dictionary-like information
- large scale results only demonstrated with one-to-one mappings, so far
- unclear how well results transfer to different classes of text

An example of acquisition: group nouns and pair nouns

Number agreement for English nouns: *group nouns* (e.g. *faculty*, *team*) are one class of exceptions to standard pattern, *pair nouns* (*scissors*, *trousers*, *binoculars*) are another. Need to know about these exceptions for monolingual NLP and MT:

- Especially in BrE, many group nouns take sg or pl agreement
- In both AmE and BrE, pl anaphors are used to refer to sg group nouns
- Pair nouns require *pair of* as a classifier (**one scissors*, ?*three scissors*)
- Translation mismatches

Acquisition of group and pair nouns

Hand coding
Very inefficient unless done semi-automatically because vast majority of nouns are regular.
Existing resources
Thesaurus, WordNet useful for candidate group and pair nouns (semantic classes).
MRDs

- Pair nouns marked as obligatory plural in LDOCE (but not unique).
- Group nouns marked in LDOCE syntax codes, but very inconsistently.
- Indirectly derive list of group nouns by looking at definition patterns (e.g. *group of* with human denoting noun).

  band: a group of musicians . . .

Tagged corpora
Patterns such as sg noun followed by pl verb (filter for conjunctions, etc)

Rotherham_NP corporation_NN are_BER

Summary of acquisition

- Hand-coding, reuse, MRDs, corpora all have their advantages and disadvantages.
- Practical acquisition involves combination of techniques, making heavy use of machine-aided and semi-automatic methods.
- Some work combines MRDs and corpora: e.g. Klavans and Tzoukermann, 1996.
- Test corpus is very important to measure progress, even if not used in any other way.
- See Lonsdale et al (1995): case study of techniques for CMU KANT system. Note: domain-specific task, but still 60,000 word and phrase vocabulary.

PART III: A case study

Project aim: build a server-based tool which allows monolingual user to construct Web site in multiple languages.

1. Project in feasibility study/proposal stage
2. Multi-site collaboration, translation to/from English
3. CSLI cogeneration approach: combines free text input, flexible templates, statistical information
4. User queried if necessary to resolve ambiguities etc using example sentences
5. Advantages of server approach:
    - Data can be continually updated
    - Very large databases feasible

An example template

I work on

Project aim:				
URL:				
Slogan:				
Time (from):	Month		Year	
(to):	Month		Year	

Instantiated template

I work on

Project name:	Archimedes			
URL:	`http://www-csli.stanford.edu/...`			
Project aim:	development of tools for disabled people			
Time (from):	Month	7	Year	95
(to):	Month		Year	

Since July 1995, I have worked on the Archimedes project —
development of tools for disabled people.

Seit Juli 1995 arbeite ich am Projekt Archimedes —
Entwicklung von Hilfsmitteln für behinderte Menschen.

Templates

ENG: work(e,x),on(e,y),pro(x,1sg),project(y)

GER: pro(x,1sg),an(e,y),arbeiten(e,x),projekt(y)

Project name: y
Project aim: y
Time: e

- Semantically fixed (with uninstantiated elements), but syntactically variable.
- Paired semantic structures for source and target language
- Translations of templates done manually
- Slots point to specific uninstantiated elements in semantics

Template lexicons

- Monolingual only: no need for bilingual pairing since translation is fixed
- Primarily to support generation
- Limited vocabulary
- Complex morphology and syntax (HPSG)
- Hand-coded, with aid of dictionaries, existing lexicons, COMLEX etc

Free-text lexicons

Monolingual entry:
`people, N pl, [WordNet node]`
Bilexicon entry:
`people_1(x) <-> menschen_1(x)`
Collocations:
`< behinderte [WordNet node] >` [positive]
`< abschalten [WordNet node] >` [negative]

- Monolingual and bilingual
- Very large coverage
- Translation by partial parsing/pattern matching — backing off to word-for-word translation
- Limited morphology and syntax: inflections, POS, subcat frames
- Semantic classification, statistical information and phrases to improve quality

Acquiring free-text lexicons

English monolingual lexicon: available resources probably adequate for first attempt: Penn morphology, tagged corpora, proper name lists etc, WordNet
Main deficiency: semantic coding with frequencies

Bilingual lexicons: depends on language:

1. Hand-coding (only for high frequency uses)
2. Reusing resources (depends on IPR)
3. Corpora
   (a) existing parallel and aligned corpora
   (b) use existing Web pages, align using HTML tags
   (c) manually create representative corpora
   (d) data about frequencies etc from running system

Concluding questions

- Is lexical acquisition really a bottleneck for successful MT? If not, what is?

  Providing syntax and basic semantics for most common usages of 100,000 words is not a big task, compared with many software development projects. Problems are missing usages, consistency, ambiguity explosion, non-compositionality.

- Can statistical techniques avoid the need for hand-compiled lexicons and knowledge bases?

- Can statistical techniques effectively address the ambiguity problem?