MODL5003 Principles and applications of machine translation
Lecture 28/04/2003

Alternatives to rule-based MT: statistical MT (SMT); example-based MT (EBMT); multilingual natural language generation (MNLG).

1. Overview

- Classification of approaches to MT
- Limitations of rule-based methods. Data-driven methods in Speech and Language Technology (SLT)
- Parallel corpora and issues of automatic alignment
- Example Based Machine Translation: metaphor of automatic translation memory and perspectives
- Statistical Machine Translation: early experiments and integration of linguistic knowledge
- Multilingual Natural Language Generation: achieving high quality in well-structured subject domains

2. Classification of approaches to MT

- Rule-based approaches (lecture 10/03/2003)

- *Direct MT*
- *Transfer MT*
- *Interlingua MT*

use formal models of our knowledge of language
("to explicate human knowledge used for translation, put it into an "Expert System")
problems:
expensive to build;
require precise knowledge, which might be not available

- Data-driven approaches (lecture today: 28/4/2003)

- *Example-based MT*
- *Statistical MT*

use machine learning techniques on large collections of available texts;
e.g. "parallel texts" (aligned sentence by sentence; phrase by phrase)
("to let the data speak for themselves")
problems:
language data are sparse (difficult to achieve saturation)
high-quality linguistic resources are also expensive

-Corpus-based support for rule-based approaches

- Multilingual Natural Language Generation

3. Limitations of rule-based methods. Data-driven methods in Speech and Language Technology (SLT)

- The cost of development of rule-based systems is too high
- Lack of adequate models (knowledge) of certain linguistic phenomena (monolingual and contrastive).
E.g., open issues in theoretical linguistics and translation studies:

- The nature of *aspect* in Slavonic languages and the ways of translating it, e.g., translations of Ukrainian:

Він читав книжку Vin chytav knyzhku he read$_{\text{(PST.IMPERF)}}$ book$_{\text{(ACC)}}$	He was reading a book
Він прочитав книжку Vin prochytav knyzhku he read$_{\text{(PST.PERF)}}$ book$_{\text{(ACC)}}$	He read (finished reading) a book

But there is no direct mapping to English aspect / tense, so lexical means should be often used for translating aspectual differences:

Нехай він читає Nexaj vin chytaje let he reads$_{\text{(NON-PAST.IMPERF)}}$	*Let him read*
Нехай він прочитає відповідь Nexaj vin prochytaje vidpovid' let he read$_{\text{(NON-PAST.PERF)}}$ answer	*Have him read the answer*

- Ways of translating definite and indefinite articles in Germanic / Romance languages into Slavonic languages using different types of word order, e.g., Russian:

The woman came out of the house	Женщина вышла из д**о**ма Zhenshchina vyshla iz doma Woman came-out of house$_{\text{(GEN)}}$
A woman came out of the house	Из дома вышла женщина Iz doma vyshla zhenshchina Of house$_{\text{(GEN)}}$ came-out woman
The woman came out of her house	Женщина вышла **и**з дому Zhenshchina vyshla íz domu Woman came-out of house$_{\text{(GEN-2)}}$

Alternative: ***data-driven methods.***
Principle: using existing translations as a prime source of information for the production of new ones (Kay, 1997, HLT survey, p. 248)

Large amounts of data contain essential knowledge for making a functional system.

- Inspired by research in Automatic Speech Recognition (ASR); applicable in other fields of SLT

- Data-driven models proved to be efficient because of:

- The availability of large amount of data and powerful computers for storing / retrieving / processing them

- Data-driven models can rectify the lack of explicit linguistic knowledge: large corpora of data contain essential information for creating functional SLT systems; the knowledge can be retrieved and used automatically

E.g.: ways of translating English word ***not*** into French with relative frequencies of translations in a parallel corpus (Hutchins, Somers, 1992, p. 321)

English	**French**
not	***ne** (0.460)**…** **pas** (0.469)*
	***ne** (0.460)**…** **plus** (0.002)*
	***ne** (0.460)**…** **jamais** (0.002)*
	***non** (0.024)*
	***pas du tout** (0.003)*
	***faux** (0.003)*

- Many machine-learning algorithms are language-independent
- Data-driven approaches allow us:
  - to account for typical phenomena systematically
  - to compare productivity of different structures in texts from different subject domains / genres.

4. Parallel corpora and issues of automatic alignment

Main sources of data for data-driven MT:

- **Parallel corpora** (collections of originals and human translations into a target language)

  Parallel corpora are richer in translation equivalents, require less sophisticated processing, but are more difficult to get.
- **Comparable corpora** (collections of texts in two different languages in the same subject domain, written around the same time)

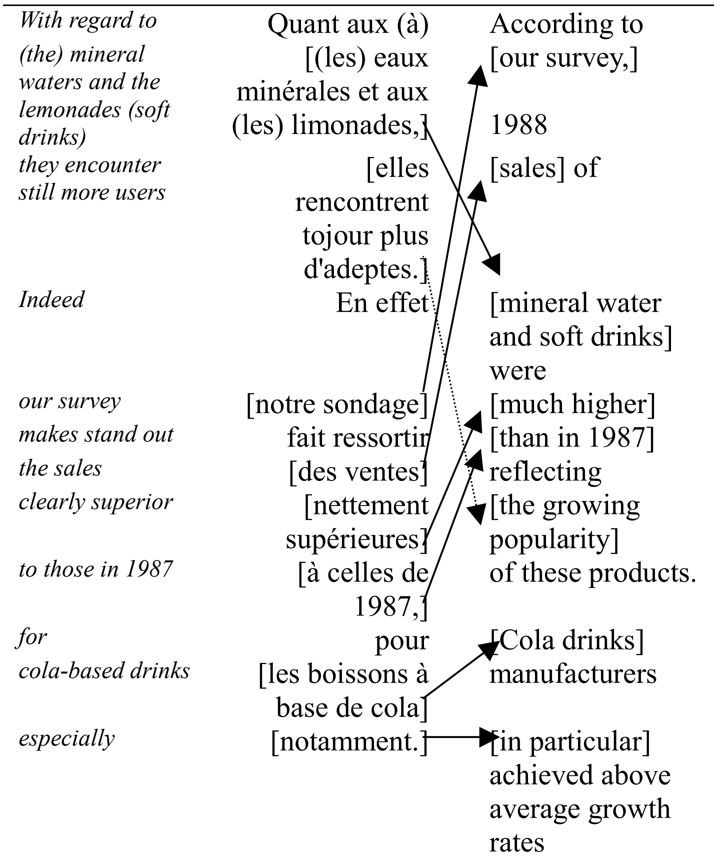
  Comparable corpora can be larger, but translation equivalents are sparse there and more difficult to identify. More research needs to be done on using comparable corpora for data-driven MT.

The goal of processing parallel / comparable corpora:

- retrieving and using translation equivalents automatically in EBMT and statistical MT
- creating lexical resources, such as bilingual dictionaries and parallel grammars to improve the quality of rule-based MT

Parallel corpora need to be aligned on the *sentence* level, on the *phrase* level and on the *word* level (when possible).

Need to know which sentences / phrases / words in texts are translations of each other



With regard to (the) mineral waters and the lemonades (soft drinks)	Quant aux (à) [(les) eaux minérales et aux (les) limonades,]	According to [our survey,] 1988
they encounter still more users	[elles rencontrent tojour plus d'adeptes.]	[sales] of
Indeed	En effet	[mineral water and soft drinks] were
our survey	[notre sondage]	[much higher]
makes stand out	fait ressortir	[than in 1987]
the sales	[des ventes]	reflecting
clearly superior	[nettement supérieures]	[the growing popularity]
to those in 1987	[à celles de 1987,]	of these products.
for	pour	[Cola drinks]
cola-based drinks	[les boissons à base de cola]	manufacturers
especially	[notamment.]	[in particular] achieved above average growth rates

C.Manning and H.Schütze, 1999, p. 469

Can the links between phrases in the figure above be made automatically?

Sentence level alignment: 90% of sentences have 1:1 alignment; the rest: 1:2; 2:1; 1:3; 3:1, etc. The example above is 2:2 alignment: much of the content of the second French sentence occurs in the first English sentence.

Sometimes the order of sentences is changed in the translation (*crossing dependencies*). These cases are difficult to align automatically with acceptable accuracy.

Summary of *sentence alignment techniques*:

- length-based alignment (Gale and Church, 1993)
- cognates (Church, 1993)
- lexical methods (Kay and Röscheisen, 1993)

word alignment techniques:

- association measures (Church and Gale, 1991), e.g.

No of aligned sentences:	cow	^ cow
vache	59	6
^ vache	8	570930

(measuring difference between the observed and expected values)

- iterative sentence-word alignment (re-computing word alignment based on its results for sentence alignment for the corpus) (Brown et al., 1990)

Problems of retrieving translation equivalents from aligned parallel corpora:

- non-literal translation: low level alignment is not possible. Solution: measuring the degree of how "literate" is the translation (distance measures: what is the level of translation equivalents).
- disambiguation information is often outside aligned items: "wearing" (clothes) can be translated as 5 different words in Japanese, so the context needs to be taken into account

Different ways of retrieving and using translation equivalents:

EBMT, Statistical MT

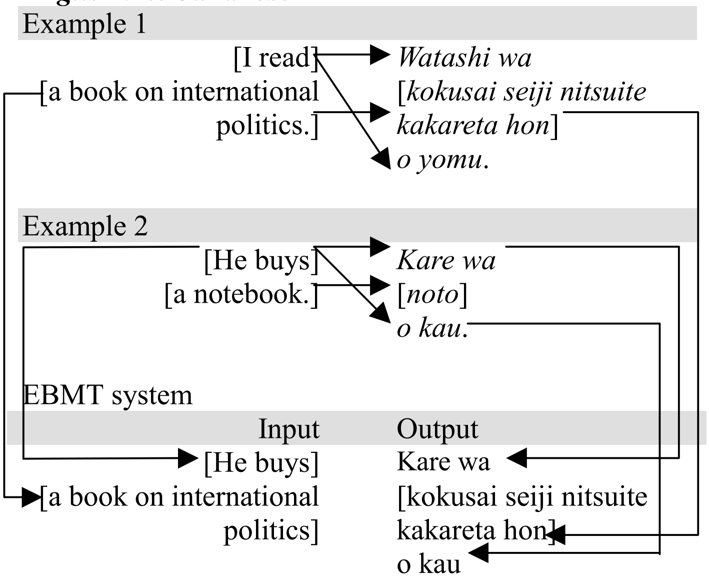
5. Example-based MT (EBMT)

EBMT (Sato & Nagao 1990), 3 stages: (Example quoted by Somers, lecture at Leeds, 2003)

- identify corresponding translation fragments (**align**)
- retrieval: **match** fragments against example database
- adaptation: **recombine** fragment into target text

(Translation Memory can be viewed as a specific case of EBMT without the adaptation stage)

English into Jananese EBMT



Example 1	
[I read]	*Watashi wa*
[a book on international politics.]	[*kokusai seiji nitsuite kakareta hon*]
	o yomu.
Example 2	
[He buys]	*Kare wa*
[a notebook.]	[*noto*]
	o kau.
EBMT system	
Input	Output
[He buys]	Kare wa
[a book on international politics]	[kokusai seiji nitsuite kakareta hon]
	o kau

Linguistic knowledge about word order, agreement, etc. is captured automatically from examples.

Issue: finding "safe points of example concatenation" (there are language-dependent and construction-dependent phenomena, e.g., "boundary friction"):

English into German EBMT

Example A	
The handsome boy ate his breakfast	Der schöne Junge aß sein Frühstück
Example B	
I saw the handsome boy	Ich sah den schönen Jungen
EBMT system	
The handsome boy entered the room	<choice: exA | exB?>

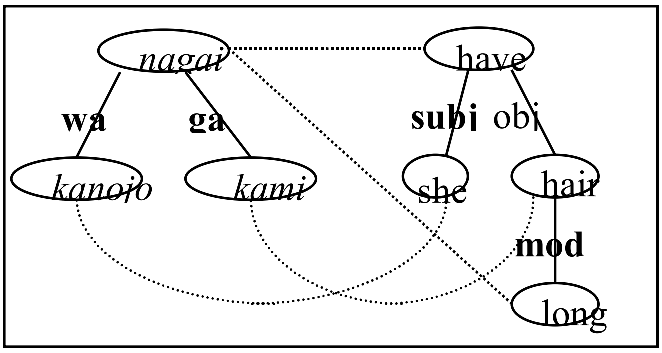
Conclusion: EBMT needs to rely on knowledge-intensive tools to combine example fragments (McTait et al., 1999)

EBMT is often integrated into traditional MT. It can be integrated into any type of rule-based systems: direct, transfer or interlingua.

Open issues in EBMT:

A. Representation and Retrieval:

- Granularity of examples: "the longer the passages, the lower the probability of a complete match, the shorter the passages, the greater the probability of ambiguity and… boundary friction or incorrect chunking" (Nirenburg et al., 1993: 48)
- Complexity of storing formats: strings, part-of-speech annotation, regular expression patterns, structures with multi-level annotation, trees…



Kanojo wa kami ga nagai.
SHE (topic) HAIR (subj) IS-LONG.
She has long hair.
(example by Somers, lecture 2003)

- Storing similar examples as a single **generalised** example (this resembles traditional transfer rules). Discovering generalised patterns automatically.

E.g.: (Brown, 1999)

John Miller flew to Frankfurt on December 3rd.
<1stname> <lastname> flew to <city> on <month> <ord>.
<person-m> flew to <city> on <date> .
Dr Howard Johnson flew to Ithaca on 7 April 1997.

- Finding adequate similarity functions (and weights for similar features) for stored examples and input sentences:
    - string edit distance (character-based, word-based, annotated word based)

E.g.: (Somers, lecture 2003):

The white horse is nice.
*a. The white horse**s** **are** nice. (4 characters, 2 words)*
*b. The white ho**u**se is nice. (1 character, 1 word)*
*c. The white **houses are** nice. (2 words)*

- semantic match (using thesaurus), e.g.:

Input:

When the paper tray is empty, remove it and refill it with paper of the appropriate size.

Character match:

When the_tray is empty, remove it and_fill it with the_appropriate sized paper.

Syntactic match:
When the bulb remains unlit, remove it and replace it with a new bulb
Semantic match:
You have to remove the paper tray in order to refill it when it is empty.

- partial match (matching fragments: sub-strings, chunks)
- multiple retrievals (ranking retrieved examples; taking the best bit of each of them: adaptation issues)

Open issues in EBMT:

B. Adaptation (recombination):

(Somers, EBMT as CBR): A solution retrieved from the stored case is almost never exactly the same as a new case. Therefore, there is a need of adapting the existing examples to a new input.

Problems:

1. Language is not always logical, information can be not explicitly present in examples, e.g.: Welsh:
stryd fawr – 'big street'
stryd fach – 'small street'
tŷ mawr – 'big house'
??? – 'small house': * *tŷ* ~~*mach*~~ *correct: ... bach*

General common-sense knowledge how language works is not enough in some cases. Solution: "model-guided repair": using knowledge about the domain, verifying if the proposed adaptation is legitimate (e.g., checking the output against a monolingual TL corpus).

2. Examples can involve different types of translation transformations: syntactic, semantic, pragmatic equivalence. Knowing how "literal" or "distant" is the translation from the original in examples is important, since examples can require different strategies for adaptation.

Proposal: Adaptation-guided retrieval (Collins, 1998:31)
Example retrieval can be scored on two counts:
(a) the closeness of the match between the input text and the example;
(b) the adaptability of the example on the basis of the relationship between the representations of the example and its translation:
e.g., more "literal" translations are easier to adapt: "adaptability" is measured by taking into account lexical and structural differences between aligned translations:

	French	English
	Ottawa abolira la très impopulaire taxe à la consommation sur les produits et les services (TPS), de type TVA, instaurée par les conservateurs,	Ottawa will abolish the very unpopular consumption tax on products and services (TPS), of the VAT type introduced by the Conservatives.
[and replace it by another ,"more equitable" tax]	et la remplacera par une autre taxe "plus équitable".	It will be replaced by another, "more equitable" tax.

(The second part of the example is less adaptable).
making distinction between good examples (best combination of retrievability and adaptability) and bad examples -- e.g., which are easy to retrieve but difficult to adapt or vice versa.

6. Statistical MT

Can be viewed as EBMT where establishing the correspondence of units in ST and TT (retrieval and adaptation) is done by statistical means.

Early years of MT involved statistical methods: were based on the assumption that our knowledge of language is insufficient for computational processing (still difficult to formalise).

(Cryptography metaphor for MT: noisy channel model: the noisy channel transformed English message into French. The problem is to recover what English speaker had in mind).

Warren Weaver's memorandum, July 1949 – proposals for tackling the obvious problems of ambiguity (or 'multiple meanings'), based on knowledge of cryptography, statistics, information theory, logic and language universals.

Statistical MT since 90's

- An experimental pure statistical system at IBM (Brown et al., 1990)
  Used the corpus of Canadian *Hansard* (records of parliamentary debates in French and English, 40,000 pairs of sentences, 800,000 words in each language)
  Evaluated by translating from French into English: limited vocabulary (1000 most frequent English words); 73 sentences:
  exact – 5%; exact + alternative + different – 48% (the rest – "wrong and ungrammatical")

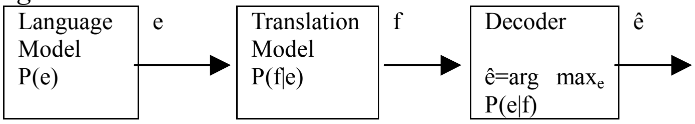
- **exact**: *Ces amendements sont certainment nécessaires*
Hansard: *These amendments are certainly necessary*
IBM: *These amendments are certainly necessary*
-**alternative**: *C'est pourtant très simple*
Hansard: *Yet it is very simple*
IBM: *It is still very simple*
- **different**: *J'ai reçu cette demande en effet*
Hansard: *Such a request was made*
IBM: *I have received this request in effect*
- **wrong**: *Permettez que je donne un exemple à la Chambre*
Hansard: *Let me give the House one example*
IBM: *Let me give an example in the House*
- **ungrammatical**: *Vous avez besoin de toute l'aide disponible*
Hansard: *You need all the help you can get*
IBM: *You need the whole benefits available*
- No prior linguistic knowledge was applied
- This experiment inspired research in SMT

Models behind the Statistical MT technology:

- Source-channel ("noisy channel") model: (Warren Weaver's "cryptography" approach).

"French sentence is viewed as "encoded" English sentence, a sentence, which was converted from English into French by some "noise" on its way to the reader. The model allows associating French and English sentences with certain numerical scores, so different "translation candidates" can be compared.

The Language Model generates an English sentence *e. P(e) – probability of e –* is an estimation of what was the sequence of words in the "underlying" English sentence (sequences that are not allowed in English will have low score) The Translation Model transmits *e* into the French sentence *f. P(f|e) – probability of a French sentence given English sentence –* is an estimation of what can be the translation of *e.* (Sequences of French words that contain words which are not translations of English words in *e* will have low scores). The decoder finds the English sentence ê which is most likely to have given rise to *f* (is not necessarily identical to *e*). *arg max – an argument with maximal value –* is the procedure of finding the highest scoring English sentence *ê* with respect to *P(e|f) – probability of an English sentence given the French sentence.*



Language model -- *P*(e): is trained on English monolingual corpus, measures how "natural", "fluent" is the resulting English sentence (Relative frequencies (probabilities) in the corpus of 2-word, 3-word… N-word sequences -- N-grams found in the output sentence are multiplied together)

Translation model -- *P*(f|e)**:** is trained on the aligned corpus, measures how "faithful", "adequate" is the resulting English sentence to the French sentence (Relative frequencies of translations of French words in parallel corpus are multiplied together).

Decoder -- arg max $_e$ **:** finds the English sentence that has the *highest value* of the multiplication of the figures provided by the two models.

Short way of writing this:

ê = arg max $_e$ *P*(e) *P* (f|e)

Problems for "pure" SMT

No notion of phrases:

to go -- aller; farmers -- les agriculteurs

Non-local dependencies:

Language models works with "fixed window" of 2, 3… N words, but more distant words can be grammatically related: E.g., 2-gram model cannot distinguish ungrammatical sentences:

What **do** you **say**?
* What do you said?
What **have** you **said**?
* What have you say?

Morphology:

Morphologically related words are treated as separate symbols, e.g., each of the 39 forms of the French verb *diriger* can be translated as *to conduct* and *to direct* -- this has to be learned separately for each form

Sparse data problem: Estimation of translation probabilities for rare words is unreliable. How to derive good characterisation of infrequent words automatically?

Later research in Statistical MT paradigm: addressing some of these problems

- The full-scale Statistical MT system "Candide" developed at IBM. Incorporates some methods of rule-based systems (analysis-transfer-synthesis paradigm) and linguistic knowledge about morphology. (Berger et al., 1994)
- Projects on using comparable corpora (which is easier to obtain, and can be much larger) together with hand-crafted morphological translation dictionaries for Statistical MT (Van Eynde, Dirix).

7. Multilingual Natural Language Generation

Addresses the task of producing high-quality multilingual documents in parallel in a well-defined subject domain.

These applications are capable of producing useful "real-world" documents

- Multilingual generation tasks are typical for large companies, which have to localise their products;
- Can be useful for individual users: e.g., application for writing business letters
- Shifting attention from translating to authoring, from ergonomics of post-editing the TT to ergonomics of producing the ST

(Hartley and Paris, "Multilingual Document Production")

Linguistic intuition behind the technology:

Much in translation is ***non-compositional***: adheres to ***conventions*** accepted in the culture of TT.

Human translators in this case use high-level (semantic and pragmatic) transformations. Examples of such translation in EBMT systems equivalents are poorly adaptable (score low on adaptability scale and are often disregarded)

E.g., (Tsujii, 1997)

English: *I will go to see my GP tomorrow*

Japanese: *Watashi wa asu isha ni mite morau*

('I will ask my GP to check me tomorrow')

Complex expressions should be directly related, so they are difficult to generalise and are very sparse in parallel corpora:

[e(X) go to see e(X's) GP] ↔
[j(X) wa j(X) no isha ni mite morau]

Using a formal model of underlying knowledge in the subject domain ("*domain knowledge base*").

- In the case of software localisation the model can be linked to interface creation tools, so much of it can be generated automatically.
- Domain knowledge base is language independent; saves cost for translation (even though the cost of creation is high)
- Changes in new versions of products require changing only the domain knowledge base. New versions of all documents for all languages, based on this model will be regenerated automatically.

Using natural language generation technology: strategic and tactical text planning, sentence planning, surface generation with grammar and lexicon.

The system has control over style, expressing given / new information, etc.

Examples of using MNLG tools (Hartley and Paris, "Multilingual Document Production")

English	**French**
To schedule an appointment	Insertion d'un rendez-vous
You must display the Appointment Editor window	Il faut afficher la fenêtre Edition de Rendez-vous
1. Choose the Appointment option from the Edit menu to display the Appointment Editor window.	1. Choisir l'option Rendez-vous dans le menu Edition pour affices la fenêtre Edition de Rendez-vous
You see the Appointment Editor window.	Vous verrez la fenêtre Edition de Rendez-vous.
2. Type the description of the appointment.	2. Introduire la description du rendez-vous.
3. Choose the start time, then chose the end time	3. Choisir l'heure de début, ensuite choisir l'heure de fin.
4. Finally, click in the Insert button	4. Enfin, cliquer sur le bouton Insertion.
The appointment appears in the appointment list	Le rendez-vous apparaît dans la liste Rendez-vous.

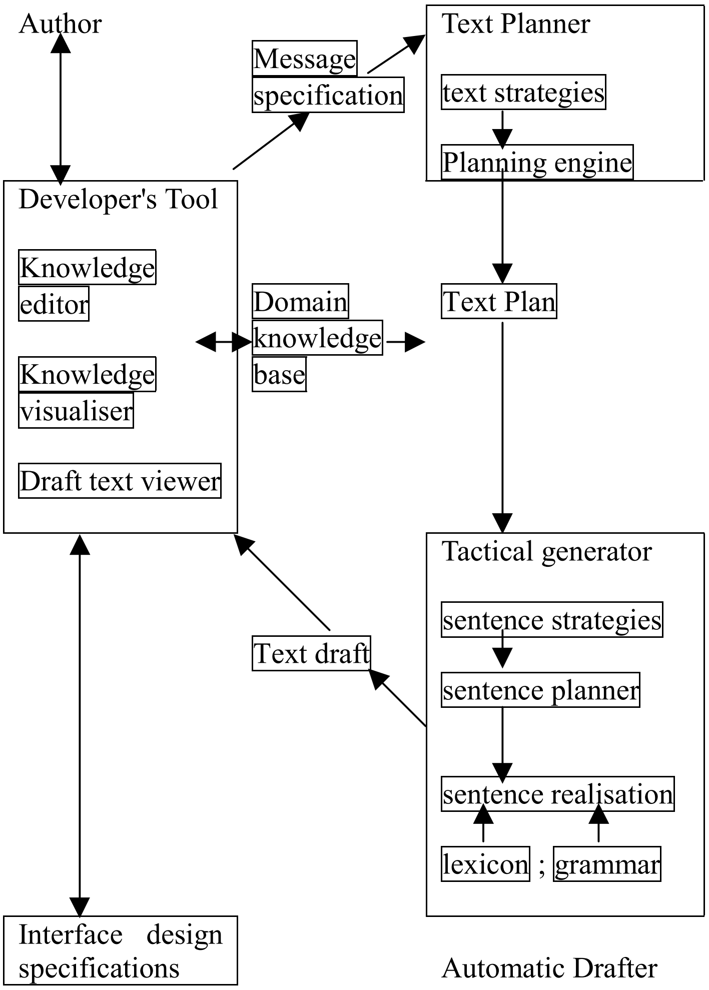
Analysis: English and French texts adhere to conventions for instruction manuals, acceptable in appropriate cultures (follow the practice of technical writers of documents in English and French corpora).

- to-infinitive clause in English vs. nominalised goal verb in French
- the user is being addressed directly (with imperative forms of verbs) in English and with a more distant form of address (with infinitive) in French.

Such variation follows naturally from the "philosophy" of MNLG, but is very difficult to model in MT.

Projects:
AGILE (University of Brighton) – "Automatic Generation of Instructions in Languages of Eastern Europe"
Non-technical overview: http://www.itri.bton.ac.uk/projects/agile/deliverables/non-tech/agile-non-technical-overview.html

Architecture of a MNLG system:


Author
Message specification
Text Planner
text strategies
Planning engine
Developer's Tool
Knowledge editor
Knowledge visualiser
Draft text viewer
Domain knowledge base
Text Plan
Tactical generator
sentence strategies
sentence planner
sentence realisation
lexicon ; grammar
Text draft
Interface design specifications
Automatic Drafter