Association for Computational Linguistics

EACL 2003

10th Conference of The European Chapter

Proceedings of the 7th International EAMT workshop on MT and other language technology tools Improving MT through other language technology tools Resources and tools for building MT

April 13th 2003

Agro Hotel, Budapest, Hungary

Association for Computational Linguistics

EACL 2003

10th Conference of The European Chapter

Proceedings of the 7th International EAMT workshop on MT and other language technology tools Improving MT through other language technology tools Resources and tools for building MT

April 13th 2003

Agro Hotel, Budapest, Hungary

The conference, the workshops and the tutorials are sponsored by:

Chief Patron of the Conference:
Dr. Ferenc Baja
Political State Secretary
Office of Government Information Technology and Civil Relations
Prime Minister's Office

Linguistic Systems BV
Leo Konst (Managing director)
Postbus 1186, 6501 BD Nijmegen, Nederland
tel: +31 24 322 63 02
fax: +31 24 324 21 16
e-mail: info@euroglot.nl, leokonst@telebyte.nl,
http://www.euroglot.nl

Xerox Research Centre Europe
Irene Maxwell
6 chemin de Maupertuis
38240 Meylan, France
Tel: +33 (0)4.76.61.50.83
Fax: +33 (0)4.76.61.50.99
email: info@xrce.xerox.com
website: www.xrce.xerox.com

ATALA
Jean Veronis
Jean.Veronis@up.univ-mrs.fr
45 rue d'Ulm
75230 Paris Cedex 5, France
http://www.atala.org

ELRA/ELDA
Khalid Choukri
choukri@elda.fr
55-57 rue Brillat Savarin
75013 Paris, France
Tel: (+33 1) 43 13 33 33,
Fax: (+33 1) 43 13 33 30
http://www.elda.fr

Order copies of ACL proceedings from: Priscilla Rasmussen, Association for Computational Linguistics, 3 Landmark Center, East Stroudsburg, PA 18301 USA, Phone +1-570-476-8006, Fax +1-570-476-0860, URL http://www.acl-web.org.

INTRODUCTION:

There is an ever-growing need for tools for translation, and Europe is envisaging a particular challenge with the enlargement of the European Union which will add 10-12 new languages.

One of the important trends emerging from the 2001 MT Summit in Santiago de Compostela was that MT is going more and more towards an integration or combination with other tools.

In the planning of this workshop we wanted to follow up on these two trends by focussing on how MT and other language technology tools can be combined in order to produce translation faster and better, and secondly on how language technology tools can support faster production of MT systems.

Various ways of combining 'proper' MT with other types of language technology tools in order to improve performance and efficiency of translation have been discussed. This may include pre-editing tools, taggers, post-editing tools, access to bilingual concordances, term extraction tools, categorisation tools, semantic clustering etc. In the programme we see a selection of such proposals for new directions for improving MT by embedding it in an environment of other tools.

Embedding is a very popular term, and it is worth noticing that we are using the concept in a slight different way here. Embedded MT normally refers to MT embed ded in another application, e.g. information retrieval, document production. But the embedding discussed here is an embedding or an enrichment of MT by combining it with other tools, namely tools which may improve the quality, even if the tools themselves are solving problems that are much simpler than the MT problem.

The other theme for the workshop was chosen because the growing demand for MT for new languages and new language pairs makes it necessary to find ways of suppor ting the production of new language pairs. Both providers and researchers are using resources, statistics, and language technology to make progress in this field. Unfortunately, this theme is not strongly represented in the papers we received.

This is the first time an EAMT workshop is organised at EACL and it has been a very good experience which can certainly be repeated in the future.

Finally, I should like to thank my collaborators in the Programme Committee: Gáb or Prószéky, Jörg Schütz and Harold Somers for their contributions and support.

Copenhagen, March 2003
Bente Maegaard, programme chair

SPONSOR:

European Association for Machine Translation (EAMT)

WORKSHOP PROGRAMME COMMITTEE:

Bente Maegaard
Center for Sprogteknologi
Njalsgade 80, DK-2300 Copenhagen, Denmark
bente@cst.dk

Gábor Prószéky
Morphologic
MorphoLogic
1118 Késmárki u. 8, Budapest, Hungary
proszeky@morphologic.hu

Jörg Schütz
IAI
Universität des Saarlandes,
Martin-Luther-Str. 14, D-66111 Saarbrcken, Germany
joerg@iai.uni-sb.de

Harold Somers
Centre for Computational Linguistics
UMIST
PO Box 88 Manchester M60 1QD, UK
Harold.Somers@umist.ac.uk

CONFERENCE PROGRAM

Sunday, April 13

9:30-9:45	Welcome and introduction Bente Maegaard (Center for Sprogteknologi, CST)
9:45-10:30	*Improving Machine Translation Quality with Automatic Named Entity Recognition* Bogdan Babych, Anthony Hartley (University of Leeds)
10:30-11:00	Coffe Break
11:00-11:45	*Two Approaches to Aspect Assignment in an English-Polish Machine Translation System* Anna Kupsc (Polish Academy of Sciences & Carnegie Mellon University)
11:45-12:30	*Multi-language Machine Translation through Interactive Document Normalization* Aurélien Max (GETA & Xerox Research Centre Europe)
12:30-14:00	Lunch
14:00-14:45	*Computer-based Support for Patients with Limited English* Harold Somers (UMIST), Hermione Lovel (University of Manchester)
14:45-15:30	Panel discussion on improving MT
15:30-16:00	Coffee Break
16:00-16:45	*Parallel Corpora Segmentation Using Anchor Words* Francisco Nevado, Francisco Casacuberta, Enrique Vidal (Universidad Politécnica de Valencia)
16:45-17:30	*An Evaluation of a Lexicographer's Workbench: building lexicons for Machine Translation* Rob Koeling (University of Sussex), Adam Kilgarriff, David Tugwell, Roger Evans (University of Brighton)
17:30-18:00	Panel discussion on resources and tools for building MT

Table of Contents

Improving Machine Translation Quality with Automatic Named Entity Recognition

Bogdan Babych
Centre for Translation Studies
University of Leeds, UK
Department of Computer Science
University of Sheffield, UK
bogdan@comp.leeds.ac.uk

Anthony Hartley
Centre for Translation Studies
University of Leeds, UK
a.hartley@leeds.ac.uk

Abstract

Named entities create serious problems for state-of-the-art commercial machine translation (MT) systems and often cause translation failures beyond the local context, affecting both the overall morphosyntactic well-formedness of sentences and word sense disambiguation in the source text. We report on the results of an experiment in which MT input was processed using output from the named entity recognition module of Sheffield's GATE information extraction (IE) system. The gain in MT quality indicates that specific components of IE technology could boost the performance of current MT systems.

1. Introduction

Correct identification of named entities (NEs) is an important problem for machine translation (MT) research and for the development of commercial MT systems. In the first place, translation of proper names often requires different approaches and methods than translation of other types of words (Newmark, 1982: 70-83). Mistakenly translating NEs as common nouns often leads to incomprehensibility or necessitates extensive post-editing. In many cases failure to correctly identify NEs has an effect not only on a local and immediate context, but also on the global syntactic and lexical structure of the translation, since proper segmentation of a source text might be seriously distorted.

However, the developers of commercial MT systems often pay insufficient attention to correct automatic identification of certain types of NE, e.g., organisation names. This is due partly to the greater complexity of this problem (the set of proper names is open and highly dynamic), and partly to the lack of time and other development resources.

On the other hand, the problem of correct identification of NE is specifically addressed and benchmarked by the developers of Information Extraction (IE) systems, such as the GATE system, created at the University of Sheffield and distributed under GPL (Cunningham et al., 1996, 2002). The quality of automatic NE identification has been evaluated at several message-understanding conferences (MUC) sponsored by DARPA. Accuracy scores for leading systems are relatively high (in comparison to other IE tasks, such as co-reference resolution, template element filling or scenario template filling). The default settings of NE recognition module of the GATE system produces between 80-90% Precision & Recall on news texts (Cunningham et al., 2002).

In this paper we describe the effect of using the GATE NE recognition module as a pre-processor for commercial state-of-the-art MT systems. The idea of our experiment is that high-quality automatic NE recognition, produced by GATE, could be used to create do-not-translate (DNT) lists of organisation names, a specific type of NE which in human translation practice is often left untranslated. (Newmark, 1982: 70-83).

In our experiment we systematically analysed the effect of incorrect NE recognition on the surrounding lexical and morphosyntactic context in MT output. We tried to establish how far NE recognition (specifically recognition of organisation names) influences grammatical well-formedness and word sense choices in the context of NEs. We compared the baseline translations (produced without NE DNT-processing) with translations produced using DNT lists (created with the GATE-1 NE recognition system), by systematically scoring cases of improvement and decline in lexical and morphosyntactic well-formedness. Texts with NE DNT-processing showed consistent improvement for all systems in comparison with baseline translations. The improvement was not lower than 20%.

This indicates that combining present-day MT systems with specific IE modules (where certain NLP problems are treated systematically) has beneficial effect on the overall MT quality.

2. Problems of NEs for MT

NEs usually require different approaches to translation than do other types of words. For example, foreign person names in Russian should be transcribed and written in Cyrillic; names that coincide with common nouns should not be looked up in the general dictionary. In some cases NEs (mostly organisation names) are not translated and preserve Roman orthography within Russian Cyrillic text. For example, in a 1000-word selection of 4 articles about the international economy on the Russian BBC World Service site, Roman-script NEs within the Cyrillic text covered 6% of the selection. The following NEs were neither translated, nor transliterated into Cyrillic: 'Nestle' (9 occurrences), 'AOL' (8); 'Buffalo Grill' (7); 'Burger King' (7); 'Diageo' (7); 'Schweisfurth (Group)' (2). In general, the practice not to translate organisation names is very common for translations into Slavic languages.

Mistakes related to the failure to distinguish between common nouns and proper nouns in MT can be very serious. For example, in our experiments an MT system translated the person name *Ray* as *Луч* ('beam of light'). Translating parts of compound NEs is also detrimental to MT quality, since it often involves incorrect segmentation of NEs: *American Telephone and Telegraph Corp.* was translated as *Американский Телефон и Компания Телеграфа* ('an American telephone and a company of a telegraph'). Yet another problem for MT systems is that failure to recognise NEs often has a negative effect on well-formedness of morphosyntactic and lexical context beyond the NEs themselves. Certain morphological features of neighbouring and syntactically related words, word order, a choice of word senses in MT output could be distorted if a NE is not correctly recognised. For example, an English phrase (1) was translated into Russian as (2):

(1) **Original**: Eastern Airlines executives notified union leaders …

(2) **MT output**: *Восточные исполнители Авиалиний уведомили профсоюзных руководителей ...*
('Oriental executives of the Airlines notified …')

This happened because the failure to identify *Eastern Airlines* as a NE led to incorrect syntactic segmentation of the sentence.

However, current MT systems allow the processing of MT input with DNT lists. Making a DNT of organisation names from the text in most cases improves not only the acceptability of NE translation, but also the overall well-formedness of the morphosyntactic and lexical context. For example, after the string *Eastern Airlines* was entered into a DNT list for the English-Russian MT system, the translation of (1) was morphologically and syntactically well-formed:

(3) **DNT-processed MT output**: *Исполнители Eastern Airlines уведомили профсоюзных руководителей ...*

Creating DNT lists manually requires much effort from the user of an MT system. However, the high accuracy in NE tagging of current IE systems, including GATE, means that DNT lists for MT can be created automatically.

The performance results reported here are based entirely on automatically created DNT lists used to process NEs.

3. Description of the experiment

In order to measure the effect of NE recognition on MT quality, we took 30 texts (news articles) from the DARPA MUC-6 evaluation set. These texts were selected because they are relatively rich in NEs, and because clean NE annotation is available for them. We used the following linguistic resources of the Sheffield NLP group:

- DARPA 'keys' – texts manually annotated with NEs;
- GATE 'responses' – the output of the automatic NE annotation of the GATE-1 system, which participated in MUC-6.

Table 1 summarises statistical parameters of this corpus. The table indicates how frequently NEs (organisation names) occur and shows that GATE 'response' figures are very close to the DARPA "key" figures.

Number of:	*For the corpus*	*Av. per doc.*	*Av. per para.*	*Av. per sent.*
Paragraphs	283	9.4	–	–
Sentences	565	18.8	2.0	–
Word occurrences	11975	399.2	42.3	21.2
Different words	3944	235.7	36.3	19.7
NE occurrences keys/ ***GATE***	544/ **510**	18.1/ **17.0**	1.9/ **1.8**	1.0/ **0.9**
Different NEs: keys/ ***GATE***	201/ **174**	7.6/ **6.7**	1.5/ **1.4**	0.9/ **0.8**

Table 1: Statistical parameters of the corpus

The density of NEs in the DARPA corpus is also characterised by Table 2:

	Manual keys	*GATE*
Paragraphs with NEs	228 (80.6%)	**218 (77.0%)**
Sentences with NEs	329 (58.2%)	**315 (55.8%)**

Table 2: NE density in the corpus

The accuracy of GATE-1 in the NE recognition task at MUC-6 (Recall – 84%, Precision – 94%, Precision & Recall – 89.06 % (Gaizauskas et al., 1995)) is such that we used the GATE output for our MT experiment, rather than the cleaner manually annotated data. Moreover, the advantage of using automatic NE recognition is that the results of the experiment should be consistent with the results for other corpora on which the NE recognition task has been performed.

Having automatically generated DNT lists of organisation names from GATE 'response' annotation, we translated the texts using three commercial MT systems:

- English-Russian 'ProMT 98' v4.0, released in 1998 (Softissimo)
- English-French 'ProMT', (*Reverso*) v5.01, released in 2001 (Softissimo)
- English-French 'Systran Professional Premium' v3.0b, released in 2000 (Systran)

Two translations were generated by each MT system:

- **a baseline translation** without a DNT list
- **a DNT-processed translation** with the automatically created DNT list of organisation names

The baseline translations were then compared with DNT-processed translations, with respect to the morphosyntactic well-formedness of the context surrounding the NEs.

3.1. Segmentation

To speed-up the process of finding contextual differences, we developed automatic tools, which allowed us to make a formal distinction between NE-internal and NE-external issues in MT. Whereas Al-Onaizan and Knight (2002) focus on the former issue, our primary interest is in NE-external differences in context caused by improved NE recognition after DNT-processing. Thus, we automatically selected paragraphs with contextual differences and highlighted different strings in these paragraphs.

The example below illustrates the output of these annotation tools:

- Different strings found in two translations are indicated by '---->'
- 'ORI' indicates the original English string in the DARPA corpus;
- 'TWS' (baseline translation) indicates a String Translated Without the do-not-translate list;

- ‘TDS’ (DNT-processed translation) indicates a String Translated with Do-not-translate list.

---->40;TDSnotInTWS: 40# Отдельно, в его регистрации
---->40;TDSnotInTWS: раскрыл детали его планов финансирования приобретения

40;ORI=40#<s> Separately, in its <ENAMEX TYPE="ORGANIZATION">SEC</ENAMEX> filing, <ENAMEX TYPE="ORGANIZATION">USAir</ENAMEX> disclosed details of its plans for financing the <ENAMEX TYPE="ORGANIZATION">Piedmont</ENAMEX> acquisition.

40;TWS= 40# Отдельно, в ее регистрации СЕКУНДЫ, USAir раскрытые детали ее планов финансирования Предгорного приобретения.

40;TDS= **40# Отдельно, в его регистрации** SEC, USAir **раскрыл детали его планов финансирования приобретения** Piedmont.

Since the amount of manual annotation was relatively small, no complex alignment for the two translated texts was implemented. Instead, we implemented a simple segmentation algorithm for paragraphs, using NE annotation in the corpus.

The segmentation was done in two stages. First, tagged NEs from the ‘ORI’ paragraph were identified and searched for in the ‘TDS’ paragraph. Then they were used as separators for the TDS: parts of the TDS between (untranslated) NEs were identified and searched for in the ‘TWS’ paragraph. If any sub-string was not found in TWS, it was printed and also highlighted in bold in TDS. This shows that strings in the context of the NE are different in the DNT-processed translation and in the baseline translation. This difference was then manually scored.

3.2. Scoring

Contextual differences between the baseline translation and the DNT-processed translation were manually scored using the scale in Table 3.

The terms ‘well-formed’ and ‘not well-formed’ refer to the local morphosyntactic or lexical context within a segment where differences occur. It remains possible that well-formed structures require post-editing at a higher level in the translated text.

The term ‘features’ refers to morphosyntactic or lexical features of certain words in the context of the NE. By ‘more correct’, we mean that the features considered in the context are correct, but the corresponding features in the compared text are wrong.

Score	*Baseline translation*	*DNT-processed translation*
+ 1	not well-formed	well-formed
+ 0.5	not well-formed;	not well-formed; some features are more correct
= 0	equally (not) well-formed	
– 0.5	not well-formed; some features are more correct	not well-formed
– 1	well-formed	not well-formed

Table 3: Scoring scheme

Here are some example strings to illustrate each score:

+1	**Original:** (It) represents 4,400 ***Western Union*** **employees** around the country.
	Baseline translation: (Он) представляет 4,400 **Западных служащих Союза** по всей стране. ('It represents 4,400 **Western employees of the Union** around the country')
	DNT-processed translation: (Он) представляет 4,400 **служащих Western Union** по всей стране. ('(It) represents 4,400 **employees of Western Union** around the country')

+0.5	**Original:** Western Union Corp. **said its subsidiary**, Western Union Telegraph Co.…
	Baseline translation: Западная Корпорация Союза **сказала ее вспомогательную,** Западную Компанию Телеграфа Союза… ('Western Corporation of a Union said **its auxiliary (case.acc.)**, Western Company of Telegraph of a Union …')
	DNT-processed translation: Western Union Corp. **Сказанный его филиал**, Western Union Telegraph Co. … ('Western Union Corp. **Its branch (case.nom) is said**, Western Union Telegraph Co.…')

=0	**Original:** *American Airlines* **Calls** for Mediation
	Baseline translation: Американские Авиалинии **Призывают** К посредничеству (American Airlines **Call(num.plur.)** for Mediation)
	DNT-processed translation: American Airlines **Призывает** К посредничеству (American Airlines **Calls(num.sing.)** for Mediation)
–0.5	**Original:** *USAir* said **that** William R. Howard, chairman and chief executive of *Piedmont*, will be elected president of *USAir*
	Baseline translation: USAir сказал **тот** Уильям Р. Говард, председатель и руководитель Предгорных, будут избраны президентом USAIR *USAir* said **that (particular)** (demonstr.pron,**nom.**) William R. Howard, chairman and chief executive of piedmont people, will be elected president of *USAir*
	DNT-processed translation: USAir сказал **того** Уильяма Ра. Говард, председатель и руководитель Piedmont, будут избраны президентом USAir *USAir* said **of that (particular)** (demonstr.pron,**gen.**) William Ra. Howard, chairman and chief executive of *Piedmont*, will be elected president of *USAir*
–1	**Original:** to discuss the benefits of **combining** *TWA* and *USAir*
	Baseline translation: чтобы обсудить выгоды от **объединения** TWA и USAIR ('to discuss the benefits of **the merge (noun) (of)** *TWA* and *USAir*')
	DNT-processed translation: чтобы обсудить выгоды от **объединяющегося** TWA и USAir ('to discuss the benefits of **the combining (participle, sing.)** *TWA* and **(of)** *USAir*')

For each MT system, we scored 50 strings showing differences. Table 4 summarises the number of paragraphs with contextual differences between the baseline and DNT-processed translations.

The figures in row 2 – *Paragraphs with contextual differences* – show to what extent DNT-processing affects the NE context for each system, showing also the percentage of these paragraphs in relation to the corresponding figure in row 1. Row 3 represents the percentage of manually scored paragraphs in relation to the figure in row 2. These figures show the likely reliability of the results for manual scoring presented in the next section.

Number of:	*Original – GATE*	*MT E-R ProMT*	*MT E-F ProMT*	*MT E-F Systran*
Paras. with NE	218	225	225	239
Paras. with contextual differences		139 (61.8%)	132 (58.7%)	207 (86.6%)
Paras. manually scored		31 (22.3%)	28 (21.2%)	30 (14.5%)
Strings with differences		211	212	411
Strings scored		50 (23.7%)	50 (23.6%)	50 (12.2%)
Diff. strings per text		7.0	7.0	13.7
Diff. paras. per text		4.6	4.4	6.9

Table 4: Paragraphs with contextual differences

Note that in row 1 there is a mismatch between the number of paragraphs with NEs in the original GATE-annotated English texts (218) and in the translations produced by the three MT systems (225, 225 and 239 paragraphs with NEs). This is because the results of NE pre-processing could be submitted to the proprietary MT systems only in the form of a DNT list, which has its limitations. The most serious potential problem is over-generation: ambiguous items, which could be either NEs or common words in different contexts, are treated as NEs in *every* context, once they are written to the DNT list. For example, the word *Labour* could be either an organisation name ('the party'), a part of a larger NE, often of a type other than organisation name (*Federal Railway Labour Act*), or a common noun ('work', as in the phrase: *rise in labour costs*). As a result, in the translated corpus there are more NEs than in the original English corpus, annotated with GATE. This is reflected in the figures presented in row 1 of Table 2. Nevertheless, the difference is relatively low (less then 10% for the worst case). Given that there are (on average) only about 2 NE occurrences per paragraph in the corpus, over-generation does not greatly affect our evaluation results.

The segmentation method described above provided us with a clear formal distinction between NE-internal and NE-external problems for MT. However, we made one exception to this distinction: in the DNT-processed English-French, Systran often incorrectly inserts definite articles for organisation names which are present in DNT list, but does not do so in the baseline translation. Our segmentation method treats these articles as part of the morphosyntactic context of NEs, and considerably increases the contextual degradation figures for Systran. But, linguistically, it is more correct to treat French articles as inner parts of NEs. Therefore, for the evaluation of contextual changes for Systran, we ignored strings where the inserted article was the only difference. As a result, Systran showed a net contextual improvement.

4. Results of the experiment

Table 5 summarises the results of the manual annotation of 50 strings containing differences for each MT system. (There are 61 scored differences for Systran, because in some strings there was more then one morphosyntactic or lexical difference).

	ProMT 1998 E-R		*ProMT 2001 E-F*		*Systran 2000 E-F*	
Mark	*N*	*Score*	*N*	*Score*	*N*	*Score*
+1*	28 =	+28.0	23 =	+ 23.0	18 =	+ 18.0
+0.5*	2 =	+1.0	5 =	+ 2.5	24 =	+ 12.0
0*	4 =	0	7 =	0	8 =	0
–0.5*	3 =	–1.5	1 =	– 0.5	1 =	– 0.5
–1*	13 =	–13.0	14 =	– 14.0	10 =	– 10.0
∑	50	**+14.5**	50	**+ 11.0**	61	**+ 19.5**
Gain		**+29%**		**+22%**		**+32%**

Table 5: Manual annotation results

N is the number of differences, annotated with that particular score. To compute the overall score for the system we multiplied the scores by the number of strings with this particular score, and added the results. The improvement was then computed by dividing the overall score by the number of scored differences: ∑score / ∑ N.

In order to see how the resulting scores change when more data is analysed, we continued scoring the English Russian ProMT 98 system, until 100 paragraphs with differences had been annotated. The results are presented in Table 6.

	ProMT 1998 E-R	
Mark	*N*	*Score*
+1*	59 =	+59.0
+0.5*	8 =	+4.0
0*	6 =	0
–0.5*	7 =	–3.5
–1*	31 =	–31.0
∑	111	**+28.5**
Gain		**+26%**

Table 6: Results for additional E-R data

We give an example of a sentence where improvement has been achieved in the DNT-processed translation for all three MT systems on several levels: morphological, syntactic and lexical.

	Original: The agreement was reached by a coalition of four of *Pan Am*'s five unions.
E-R ProMT	**Baseline translation**: Соглашение было достигнуто коалицией четырех Кастрюли пять союзов Ама. ('The agreement was reached by a coalition of four of a Saucepan five unions of Am.')
	DNT-processed translation: Соглашение было достигнуто коалицией четырех из пяти союзов Pan Am. ('The agreement was reached by a coalition of four out of five unions of *Pan Am* ')
E-F ProMT	**Baseline translation**: L'accord a été atteint par une coalition de quatre de casserole cinq unions d'Am. ('The agreement was reached by a coalition of four of saucepan five unions of Am.')
	DNT-processed translation: L'accord a été atteint par une coalition de quatre de cinq unions de Pan Am. ('The agreement was reached by a coalition of four of five unions of Pan Am.')
E-F Systran	**Baseline translation**: L'accord a été conclu par une coalition de quatre de la casserole étais cinq syndicats. ('The agreement was reached by a coalition of four of the saucepan was five trades-unions.')
	DNT-processed translation: L'accord a été conclu par une coalition de quatre de Pan Am's cinq syndicats. ('The agreement was reached by a coalition of four of Pan Am's five trades-unions.')

Here are further typical cases of morphosyntactic improvement in the translated material:

Improved syntactic segmentation:	
	Original: Representatives for the 5,400-member *Allied Pilots Association* didn't return phone calls.
E-R ProMT	**Baseline translation**: Представители для *Союзнических Пилотов с 5,400 членами Ассоциация* не возвращали обращения по телефону. ('Representatives for the *Allied Pilots with 5,400 members Association* didn't return phone calls.')
	DNT-processed translation: Представители для Allied Pilots Association с 5,400 членами не возвращали обращения по телефону. Representatives for the *Allied Pilots Association* with 5,400-members didn't return phone calls.

Improved proper / common disambiguation:	
	Original: A spokesman for the company said *American* officials 'felt that …'
E-F ProMT	**Baseline translation**: Un porte-parole de la société a dit que les fonctionnaires *américains* 'ont estimé que …' ('A spokesman for the company said that the American [US] officials 'felt that …'')
	DNT-processed translation: Un porte-parole de la société a dit que les fonctionnaires *d'Américan* 'ont estimé que …' ('A spokesman for the company said that the officials of American 'felt that …'')

Improved word order:	
	Original: *USAir disclosed details* of its plans for financing …
E-F ProMT	**Baseline translation**: *USAir les détails révélés* de ses plans pour financer … ('USAir the details revealed (*Past participle*) of its plans for financing …')
	DNT-processed translation: *USAir a révélé les détails* de ses plans pour financer … ('USAir revealed (*Verb*) the details of its plans for financing …')

Improved lexical or syntactic disambiguation:	
	Original: TWA *stock closed* at $28 …
E-F Systran	**Baseline translation**: *Fermé courant* de TWA à $28 … ('Closed (*Past participle*) current (*Noun/Present participle*) of TWA at $28 …')
	DNT-processed translation: *L'action* de TWA *s'est fermée* à $28 … ('The stock of TWA closed (*Verb*) at $28 …')
	Original: National Mediation Board is expected to release Pan Am Corp. and its Teamsters union from their long-stalled contract negotiations.
E-R ProMT	**Baseline translation**: Национальное Правление Посредничества, как ожидается, *выпустит* Кастрюлю - Корпорация и ее союз Водителей от их долго-остановленных переговоров контракта. ('National Mediation Board is expected to release [put on the market] a Saucepan - Corporation and its union of drivers from their long-stalled contract negotiations.')
	DNT-processed translation: National Mediation Board, как ожидается, *освободит* Pan Am Corp. И его союз Teamsters от их долго-остановленных переговоров контракта. 'National Mediation Board is expected to release [make free] Pan Am Corp. and its Teamsters union from their long-stalled contract negotiations.')

5. Conclusions

The results indicate that combining IE technology with MT has a great potential for improving the state-of-the art in output quality. Taking advantage of efforts to resolve specific linguistic problems – as has happened with NE recognition within the IE framework – improves not only the treatment of that phenomenon by MT, but also morphosyntactic and lexical well-formedness more generally in the wider context of the target, thus boosting the overall quality of MT. Our results show that modern MT systems still leave room to achieve a considerable improvement. Further gains in performance may be anticipated by harnessing other focussed technologies, such as word sense disambiguation, to MT.

We noted also that the scale of the improvement for particular MT systems correlates with the baseline quality of MT: it is more difficult to achieve improvement for a system which produces high-quality well-formed structures without DNT-processing. The improvement which is possible with NE DNT-processing is lowest for the English-French ProMT (Reverso) system. This system was ranked higher than English-French Systran by human evaluators in an experiment conducted by (Rajman and Hartley, 2001) using data from DARPA's 1992-1994 series of MT evaluations (White, et al, 1994). These human evaluations confirmed the ranking predictions of an automatic evaluation algorithm which correlated the fluency, adequacy and informativeness scores awarded by human evaluators to the DARPA corpus with syntactic and semantic attributes of the corpus. In this respect, the measures of contextual improvement after DNT-processing with lists of NEs (organisation names) produced by GATE could be seen as a possible evaluation score for MT systems, which could lead to establishing a reliable quality scale for MT systems.

Future work will look at the sensitivity of the performance gain to corpus size and variation. Table 6 shows that the difference in the score for 50 annotated paragraphs and the score for 100 paragraphs for E-R ProMT98 is 3%. In general, different occurrences of the same NE tend to have a similar morphosyntactic context, so they constantly tend to either improve or worsen the quality. In a particular text, the same NEs tend to re-occur. As a result, an improvement or a decline in quality is usually not homogeneous across corpora, but is more constant for a particular text. The score changes in more or less homogeneous chunks of text. For E-R ProMT 98 MT system the average size of such chunks is about 7 differences (See Table 3, row 6 'Different strings per text'). For E-R ProMT 98, the value of each '+1' or '–1' score after 50 annotated differences is ±2%, so one text can potentially change the score by about ±14%. After checking 100 differences, the value of each '+1' or '–1' score becomes ±1%, so a new text could change the score by ±7% on average. In the case of E-R ProMT 98, scoring 50 additional new strings (about 7 new texts) changed the overall score by –3%. This indicates that, for our corpus, there is a reliable improvement after NE DNT-processing, but more work remains to be done.

Other future work will consider the well-formedness or acceptability of the NEs themselves.

References

Al-Onaizan, Y. and K. Knight. 2002. Translating Named Entities Using Monolingual and Bilingual Resources. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Philadelphia, July 2002. pp. 400-408.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).* Philadelphia, July 2002.

Cunningham, H., Y. Wilks and R. Gaizauskas. 1996. GATE -- a General Architecture for Text Engineering. *Proceedings of the 16th Conference on Computational Linguistics (COLING-96),* Copenhagen, Aug, 1996.

Gaizauskas, R., T. Wakao, K. Humphreys, H. Cunningham, Y. Wilks. 1995. University of Sheffield: Description of the LaSIE system as used for MUC-6. *Proceedings of the 6th Message Understanding Conference (MUC-6).* Morgan Kaufmann, pp. 207-220.

Newmark, P. 1982. *Approaches to translation.* Pergamon Press, Oxford, NY.

Rajman, M. and T. Hartley. 2001. Automatically predicting MT systems ranking compatible with Fluency, Adequacy and Informativeness scores. *Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit VIII.* Santiago de Compostela, September 2001. pp. 29-34.

White, J., T. O'Connell and F. O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons and future approaches. *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*. Columbia, MD, October 1994. pp. 193-205.

An Evaluation of a Lexicographer's Workbench: building lexicons for Machine Translation

Rob Koeling
COGS, University of Sussex
robk@cogs.susx.ac.uk

Adam Kilgarriff, David Tugwell, Roger Evans
ITRI, University of Brighton
{adam,david,roger}@itri.bton.ac.uk

Abstract

NLP system developers and corpus lexicographers would both benefit from a tool for finding and organizing the distinctive patterns of use of words in texts. Such a tool would be an asset for both language research and lexicon development, particularly for lexicons for Machine Translation (MT). We have developed the WASPBENCH, a tool that (1) presents a "word sketch", a summary of the corpus evidence for a word, to the lexicographer; (2) supports the lexicographer in analysing the word into its distinct meanings and (3) uses the lexicographer's analysis as the input to a state-of-the-art word sense disambiguation algorithm, the output of which is a "word expert" for the word which can then disambiguate new instances of the word. In this paper we describe a set of evaluation experiments, designed to establish whether WASPBENCH can be used to save time and improve performance in the development of a lexicon for Machine Translation or other NLP application.

1 Motivations

On the one hand, Human Language Technologies (HLT) need dictionaries, to tell them what words mean and how they behave. On the other hand, the people making dictionaries (herafter, lexicographers) need HLT, to help them identify how words behave so they can make better dictionaries. This potential for synergy exists across the range of lexical data - in the construction of headword lists, for spelling correction, phonetics, morphology and syntax, but nowhere is it truer than for semantics, and in particular the vexed question of how a word's meaning should be analysed into distinct senses. HLT needs all the help it can get from dictionaries, because it is a very hard problem to identify which meaning of a word applies, and if the dictionary does not provide both a coherent and accurate analysis of what the meanings are, and a good set of clues as to where each meaning applies, then the enterprise is doomed. The lexicographer needs all the help they can get because the analysis of meaning is the second hardest part of their job (Kilgarriff, 1998), it occupies a large share of their working hours, and it is one where, currently, they have very little to go on beyond intuition.

Synergy between HLT and lexicographer becomes a possibility with the advent of the corpus. Lexicographers have long been aware of their great need for evidence about how words behave. The pioneering project was COBUILD (Sinclair, 1987) and its first offering to the world, the Collins COBUILD English Dictionary came out in 1987. The basic working methodology, in those early days, was the 'coloured pens' method. A lexicographer who was to write an entry for a word, say *pike*, was given the corpus evidence for *pike* in the form of a key-word-in-context printout. They then read the corpus lines, identifying different meanings as they went along, assigning a colour to each meaning and marking each corpus line with the appropriate colour. Once they had marked all (or almost all - there are always anomalies) the corpus lines, they could then go back to write a definition for each sense, using, eg, the red corpus lines as the evidence for the first meaning, the green as the evidence for the second, and so on.

In this scenario, note that a meaning, or word sense, corresponds to a cluster of corpus lines. This is a representation that HLT can work with. As corpus-based HLT took off, in the 1990s, researchers such as (Gale et al., 1993) explored corpus methods for word sense disambiguation (WSD). Here the correspondence between word senses and sets of corpus lines was taken at face value, with a set of corpus lines which were known to belong to a particular sense being used as a training set. A machine-learning algorithm was then able to use the training set to induce a word expert which could decide which sense a new corpus instance belonged to.

So the stage is set for software which both uses HLT to support the corpus lexicographer in developing good meaning analyses, and uses the meaning analysis, realised as corpus evidence, to support accurate WSD. This is what the WASPBENCH aims to do.

1.1 The WASPBENCH system

Behind the current implementation of the English WASPBENCH lies a database of 70M instances of grammatical relations for English. These are 5-tuples:
$< gramrel, word1, word2, particle, pointer >$
gramrel can be any of a set of 27 core grammatical relations for English (including *subject, subject-of, object, object-of, modifier, and/or, PP-comp*), *word1* and *word2* are words of English (nouns, verbs or adjectives, lemmatized to give dictionary headword form; *word2* may be null), *particle* is a particle or preposition, so that grammatical relations involving prepositions as well as two fully lexical arguments can be captured. For all relations except *PP-comp* it is null. *Pointer* points into the corpus, so we can identify where the instance occurs and retrieve its context if required. Examples of 5-tuples are

PP-comp,look,picture,at,1004683
object, sip, beer, -, 1005678

The database was prepared by parsing a lemmatised, part-of-speech-tagged version of the British National Corpus, a 100M word corpus of recent spoken and written British English.[1]

Using this database, WASPBENCH prepares a set of lists for each *word1* in which, for each *gramrel*, the words which occur frequently and with high mutual information as *word2* are identified and sorted according to their lexicographic salience. This set of lists is presented to the lexicographer for whom it is a useful summary of the word's behaviour. This is a *word sketch* (Kilgarriff and Tugwell, 2001b).

The word sketch is a good starting point for the lexicographer to analyse the different meanings (step 1). They study it. All underlying corpus evidence is available at a mouseclick, in case they are unsure what contexts *word1* occurs in *gramrel* with *word2* in. They reach preliminary opinions about the different meanings the word has. They assign a short mnemonic label to each sense, and type the labels into a text-input box provided. Hitting the "set senses" button updates the word sketch, with each collocate now having a pull-down menu through which it can be assigned to one of the senses.

The lexicographer then spends some time – typically some thirty minutes for a moderately complicated word– assigning collocates to senses (step 2). The majority of high-salience
$< collocate, gramrel >$ pairs relate to one sense of a word only (in accordance with Yarowsky's "one sense per collocation" dictum (Yarowsky, 1993)), and it is usually immediately evident which sense is salient, so the task is not unduly taxing. The lexicographer does not have to assign all, or any particular, collocate, and any collocate which is associated with more than one sense should be left unassigned.

When the lexicographer has assigned a good range of collocates, they press "submit". The WSD algorithm takes over, using the corpus instances where the collocates assigned by the lexicographer apply as the clusters of instances corresponding to a sense, and bootstrapping further evidence about how other corpus instances are assigned (step 3). The algorithm produces a *word expert* which can disambiguate new instances of the

[1] `http://info.ox.ac.uk/bnc`

word. The algorithm currently in use is a reimplementation of Yarowski's decision list learner (Yarowsky, 1995).

1.2 WASPBENCH and Machine Translation

WASPBENCH is designed particularly with the needs of MT lexicography in mind. In that context, the components of the problem take on a slightly different form, sometimes with different names. MT has long needed many rules of the form,

> *in context* **C**, *translate source language word* **S** *as target language word* **T**

The problem has traditionally been that these rules are hard for humans to identify, and, as there is a large number of possible contexts for most words and a large number of ambiguous words, a very large number of rules is needed. In step (1), the word sketch, WASPBENCH identifies and displays to the user a good set of candidate rules but with the target word **T** unspecified. In step (2), it supports the assignment of target words, by the lexicographer, for a number of the rules. In step (3), it takes this small set of rules and uses a bootstrapping algorithm to automatically identify a very large set of rules, so the word can be appropriately translated wherever it occurs (Kilgarriff and Tugwell, 2001a).

2 Evaluating WASPBENCH

Evaluating how successful we have been in developing the WASPBENCH presents a number of challenges.

- We straddle three communities - the (largely commercial) dictionary-making world, the (largely research) Human Language Technology (and specifically, WSD) world, and the (part commercial, part research) MT world, all with very different ideas about what makes a technology useful.

- There are no precedents. WASPBENCH performs a function – corpus-based disambiguating-lexicon development with human input – which no other technology performs. We believe no other technology provides even a remotely similar combination of inputs (corpus + human) and outputs (meaning analysis + word expert). This leaves us with no other products to compare it with.

- On the lexicography front: human analysis of meaning is decidedly 'craft' (or even 'art') rather than 'science'. WASPBENCH is aiding the practitioners of this craft in doing their job better and faster. But, in the dictionary world, even qualitative analyses of the relative merits of one meaning analysis as against another are rare treats. Quantitative evaluations are unheard of.

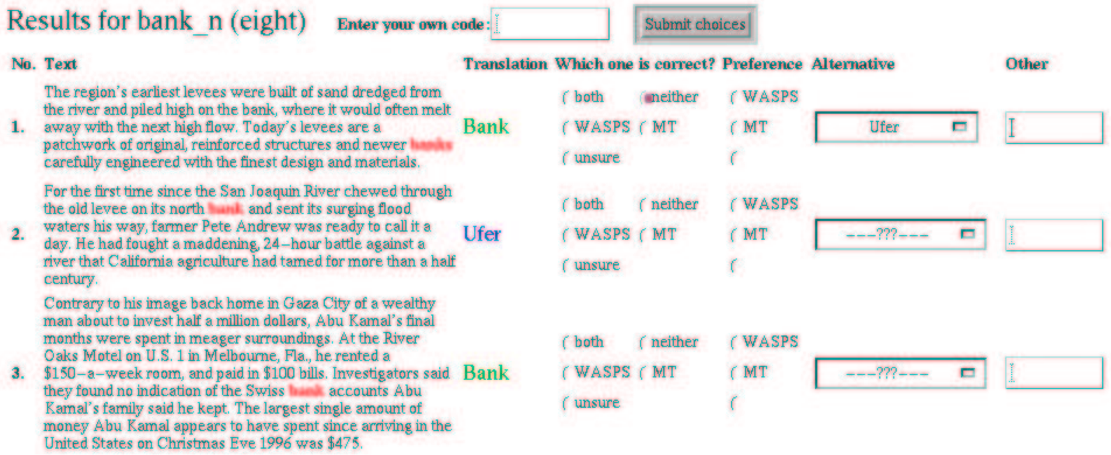
- A critical question for commercial MT would be "does it take less time to produce a word expert using WASPBENCH than using traditional methods, for the same quality of output". We are constrained in pursuing this route because we do not have access to MT companies' lexicography budgets, and moreover consider it unlikely that MT companies would view the production of disambiguation rules as a distinct function in the way that we do.

In the light of these issues, we have adopted a 'divide and rule' strategy, setting up different evaluation themes for different perspectives. We have pursued five different evaluation strategies. One of them is the subject of this paper.[2] Of the other strategies, we only mention the application of word sketches within a large scale commercial lexicography project here (the production of Macmillan English Dictionary for Advanced Learners) (Kilgarriff and Rundell, 2002). The set of experiments that we report on in this paper explored the performance of WASPBENCH-based translations in comparison with translations produced by commercial MT systems.

3 Experimental setup

A group of twelve people were involved in the experiment. All were students in translation studies at the University of Leeds. None of them had a

[2] A report bringing together evidence from all evaluation approaches is in preparation.



Results for bank_n (eight) Enter your own code: Submit choices

No.	Text	Translation	Which one is correct?	Preference	Alternative	Other
1.	The region's earliest levees were built of sand dredged from the river and piled high on the bank, where it would often melt away with the next high flow. Today's levees are a patchwork of original, reinforced structures and newer banks carefully engineered with the finest design and materials.	Bank	○ both ● neither ○ WASPS ○ MT ○ unsure	○ WASPS ○ MT ○	Ufer	
2.	For the first time since the San Joaquin River chewed through the old levee on its north bank and sent its surging flood waters his way, farmer Pete Andrew was ready to call it a day. He had fought a maddening, 24–hour battle against a river that California agriculture had tamed for more than a half century.	Ufer	○ both ○ neither ○ WASPS ○ MT ○ unsure	○ WASPS ○ MT ○	---???---	
3.	Contrary to his image back home in Gaza City of a wealthy man about to invest half a million dollars, Abu Kamal's final months were spent in meager surroundings. At the River Oaks Motel on U.S. 1 in Melbourne, Fla., he rented a $150–a–week room, and paid in $100 bills. Investigators said they found no indication of the Swiss bank accounts Abu Kamal's family said he kept. The largest single amount of money Abu Kamal appears to have spent since arriving in the United States on Christmas Eve 1996 was $475.	Bank	○ both ○ neither ○ WASPS ○ MT ○ unsure	○ WASPS ○ MT ○	---???---	

Figure 1: Snapshot of the evaluation screen

specific background in lexicography. They were all native or near-native speakers of both English and the language they worked with for the experiment. The students worked with Chinese (4), French (3),German (2) and Italian (1).[3]

We asked the participants to work with the WASPBENCH; creating word experts for the selected words. This task gave us information about how the users experienced using the workbench, either explicitly, by giving us feedback, or implicitly by supplying us with data. This part of the experiment created the word experts. The other task was to evaluate the word experts. We applied their word-experts to a set of previously unseen test sentences and compared the output of the WASPBENCH with the output of a commercial MT system.

Creating the word experts The main task for the participants was to use the WASPBENCH to create word experts for a list of selected ambiguous English words. The evaluation task focussed on translation. The user was asked to use the WASPBENCH in order to find out how the word was used in English (i.e. as represented by the BNC) and how the different uses of this word would be translated into a target language of the participant's choice. After the user has chosen the translations for the word and selected the clues giving evidence for when the word should receive a particular translation, the user submits the data and the WASPBENCH infers further rules to complete the word expert. The user is presented the rule set and can manually inspect it. If they are happy with the set, they can decide to submit the word expert and continue with the next word. If they are not happy with the rule set, they can return to the wordsketch definition form and add to or amend their input. After submitting, the word expert is applied to a set of test sentences.

Assessing the results Evaluating a word expert is like evaluating the work of a translator. The work of a translator can be judged by someone else, who can disagree on certain decisions made by the translator. The disagreement can be a matter of personal style. The assessment task here involves the same kind of problem. In this experimental paradigm we do not define beforehand what the desired translation is. Every subject may identify a different set of target translations for each word and even if they work with the same set, people might disagree on the preferred translation of a ord in a particular context. There is no gold standard and thus we cannot evaluate the decisions automatically. Therefore we asked the participants to assess the word experts' judgements.

The assessment task can best be introduced by looking at a screenshot. In figure 1 we present part of the evaluation screen with the results of ap-

[3] Two more students worked with Japanese, but at the time of the experiment we did not have the MT translations for Japanese avaliable. Their word experts were evaluated in a different way. We do not discuss these results in this paper.

plying the word expert made by participant ‘one’ for the noun *bank* to the set of 45 test sentences. The assessor is asked to enter their own number for identification purposes. The second column gives the test sentences with the word we are interested in (here *bank*) highlighted. The third column presents the word expert’s translation. The assesser is asked to judge the correctness of the translation in this particular context in the fourth column. It was our intention to either include the whole translated sentence as generated by the MT system on the screen (with the target word highlighted) or just the translated target word. However, last minute technical problems made this impossible and we had to provide the MT system output on paper. The assesser was asked to decide which translation was correct in the given context. The options given were ’WASPS’, ’MT’, ’both’, ’neither’, ’unsure’ and combinations like ’both correct, but WASPS preferable’.

In case they disagree with the translation offered, they can pick their preferred translation from the pulldown menu in the fifth column (**Alternative**). This pulldown menu offers all the other suggested target translations for *bank* as defined by participant ‘one’. In case the assesser thinks the proper target translation is not available, their choice can be entered in the last column (**Other**).

After judging all 45 test sentences, the assesser is asked to submit the form by pressing the button in the right upper corner.

3.1 Instruction and Available Time

Most participants had not worked with the WASPBENCH before. They were given a theoretical introduction and the opportunity afterwards to explore the user interface and its functionality by creating a word expert. The participants were allowed plenty of time to create the word expert and play with the WASPBENCH. They then applied the word expert to a set of test sentences and inspected the results, to conclude the introduction.

After the instruction session, approximately 4 days were allowed for working on the task: about two days for creating word experts and two days for assessment. The participants were instructed to take their time to create the word experts, but to keep in mind that we did not expect perfection. In order to finish all 33 words in two working days, only aproximately 30 minutes per word was available. We did not expect them to complete the full list. To ensure that every word on the list would be covered by equally many subjects, everyone was asked to start at a different position in the list of words.

3.2 The Data

Words For the experiment we chose a set of words that are clearly ambiguous in English. We only selected words that were fairly, but not extremely, common (i.e. with 1,500 - 20,000 instances in the BNC). A total of 33 words were selected: 16 nouns, 10 verbs and 7 adjectives. Some of the words have just two clearly distinct meanings in English, others have more. There may of course also be further, more subtle meaning distinctions. All of the words were checked to confirm that the ‘clearly distinct meanings’ receive different translations in at least one of the languages at our disposal (Dutch, German and French). While we had identified a set of meanings for the words in the course of this process, this set was never shown to the participants. They were asked to create their own word expert with its own inventory of meanings/translations. This might result in different sets of target translation for different languages. In some languages two distinct different meanings might be translated with the same word, while subtle meaning differences might produce different translations in the target language. It is, of course, possible that, where more than one participant was working on the same language, they disagreed on the one set of target translations.

Test Data In order to test the performance of the word experts, we selected for every word between 40 and 50 text fragments containing the target word. These fragments consisted of the complete sentence in which the word occurred plus one or two surrounding sentences. The test sentences were selected from the North American News Text Corpus.[4] Random samples were taken from the corpus and inspected for suitability. This was done

[4] Available from the Linguistic Data Consortium.

Language	**Wasps**		**MT**		**both**	**neither**	**unsure**
German	**0.60** %	(0.41)	**0.28** %	(0.09)	0.19 %	0.26 %	0.05 %
French	**0.61**	(0.24)	**0.45**	(0.07)	0.37	0.28	0.04
Chinese	**0.68**	(0.32)	**0.42**	(0.05)	0.37	0.23	0.03
Italian	**0.67**	(0.44)	**0.29**	(0.06)	0.23	0.22	0.05
All	**0.64**	(0.35)	**0.36**	(0.07)	0.29	0.25	0.04

Figure 2: WASPBENCH results compared with MT per language

to make sure that the samples were usable (some samples, like words from headlines, did not have much surrounding text) and to ensure that for every identified distinct meaning there were at least some test sentences available. If we had chosen a large set of test sentences from the corpus, we could have relied on pure random selection to take care of the proper meaning distribution, but a considerably larger sample than the 45 test sentences taken here would be necessary to rely on that.

The fact that we used an American news corpus for the test sentences and that the WASPBENCH currently uses the BNC for creating the word experts caused another problem: some words are used differently in British and American English, for example *lot* which has the 'parking space' meaning in American but not British English.

MT translation The MT translations were produced with BabelFish from Systran.[5] The individual fragments (i.e. the sentence wit the ambiguous word in it plus 1 or 2 surrounding sentences) were submitted as seperate paragraphs to the translation engine.

4 Evaluation of the Results

A total of 240 word experts were produced for 32 words.[6] This means that an average of 7.5 word experts per word are available. There were at least 5 different word experts for any word, the maximum number of word experts for one word is 10.

The results for the different words depend very much on the perceived ambiguity of the word and how closely related the different meanings for that word are. For example, a noun like *bank* with two clear and distinct meanings ('financial institution' and 'river bank') gave very good results, while the results for very ambiguous words like the noun *line* were quite poor. The table in figure 3 gives an overview of the results of applying the word experts to the test sentences and comparing the translation of the target word with the translation for that word given by the MT system. The data is presented here per language. The figures in bold face give the overall percentage of cases were the WASPBENCH or the MT system was considered to be right. This number is the sum of the percentage of cases were only WASPBENCH /MT was right (percentage in brackets after the bold face) and those cases where both were considered to have given the right translation.

The table in figure 3 presents the data per PoS tag. This table shows that the WASPBENCH performs slightly better on nouns (which is consistent with the comments we got from the participants, who thought that the nouns were less problematic than the verbs and adjectives).

The data shows that the WASPBENCH results consistently outperform the MT results by a considerable margin. We do have to take into acccount that the sample sentences in the test sets we used here were not taken from one particular domain, but a sample of general text. The gains for translating domain specific text might be less dramatic.

5 User Experience with the Workbench

The evaluation task did not only provide data; it also gave us feedback on working with the workbench. Many comments were given on the pre-

[5] Available over the web via Altavista: http://babelfish.altavista.com/

[6] We experienced problems with one of the nouns. The data for this word (*film*') was discarded.

PoS	Wasps		MT		both	neither	unsure
Noun	**0.69**	(0.34)	**0.40**	(0.06)	0.35	0.24	0.02
Verb	**0.61**	(0.29)	**0.38**	(0.05)	0.32	0.27	0.06
Adjective	**0.63**	(0.32)	**0.41**	(0.10)	0.31	0.24	0.04

Figure 3: WASPBENCH results compared with MT per Part of Speech

sentation of the data, missing navigation abilities, buttons and correction facilities and other user-interface issues. We will incorporate suggestions into future releases of the workbench.

An important issue is that people have difficulties with many of the grammatical relations, and instead, focus on example sentences. This is time-consuming and it would be better if we could clarify the grammatical relations, either on the same screen, or on demand (for example by making help available).

A source of confusion and irritation is PoS tagger errors and errors made in predicting the grammatical relations. It is clear that these components are critical for the usability of the workbench.

6 Conclusions and Further Research

We have already mentioned that the evaluation experiment have provided us with valuable feedback on how people experience working with the WASPBENCH, giving us the opportunity to further develop the workbench. Several changes in the user interface will be made and will improve the usability of the tool. The main objective for this particular experiment, however, was to investigate how well the word-experts created with the WASPBENCH help to disambiguate words in a translation task. These experiments show that with the WASPBENCH it is possible to create word sense disambiguation rules that help translation of ambiguous words enormously without spending a whole lot of time in creating these rules. The results show that people, with no prior experience using the workbench, are able to create disambiguation rules that outperformed a well-established MT system by a great length, even though they had limited time to spend on creating the rules and did not have the opportunity to improve on their efforts.

While thinking about the WASPBENCH as a tool for improving WSD for MT systems, one of the questions we asked ourselves was: “does it take less time to produce a word expert using WASPBENCH than using traditional methods, for the same quality of output”. Even though we can’t answer this question, we do know now that we can improve substantially upon the quality of the output. We can also estimate the cost (in time or money) to create disambiguation rules for all the words and estimate the improvement in quality it will give us.

Another important aspect of the evaluation results is the fact that the results for the different languages are very similar. We feel that consistency is important for a disambiguation tool. Even though the word experts created by the participants will always be different, they should ideally behave similarly. In another experiment (Koeling and Kilgarriff, 2002) we looked explicitly at the consistency of results by comparing word experts (same word, same target language) made by several people. In that experiment we found more evidence for our consistency claim.

Even though we feel that these experiments show that the WASPBENCH succesfully meets many of the goals we had in mind when we designed the workbench, there are still ways to improve the current system. The fronts on which we would like to develop the WASPBENCH include:

exploring alternative WSD algorithms

(Yarowsky and Florian, 2002) show that ”winner-take-all” algorithms, are sometimes preferable, but sometimes cumulative algorithms, where evidence from different clues is summed, perform better. We would like to explore how we might match the algorithm-type to the data instance.

interactivity Currently there is only minimal support for a ’second round’ of the lexicographer revising their meaning analysis according to the feedback provided by the WSD algorithm. We would like the system to enter a dialogue with the lexicographer, whereby it identified anomalies and facilitated revisions to the meaning analysis.

multiwords Although some fuctionality for multiwords is already supported, for phrasal verbs and subcategorising nouns and adjectives, through the three-argument *prep_n* relation, we would like to extend system functionality by permitting the user to input multiwords, for which collocations would be found.

thesaurus We have already produced a thesaurus from the database (see http://wasps.itri.bton.ac.uk), using Lin's similarity measure (Lin, 1998). We would like to use the thesaural classes in the word sketches and elsewhere, so that evidence from words in the same thesaural class could be pooled, and inferences drawn where two words were not encountered together, but their thesaural classes had high mutual information.

other languages Developments for a number of languages other than English are under way. Once we have two databases of grammatical relations, based on comparable corpora, for different languages, the potential for mapping tuples between the databases (using a bilingual dictionary) arises.

new corpora there's no data like more data, and both wordsketch production and the WSD learning algorithm work better, the more they are fed. Using the BNC, we have insufficient data to say much about words beyond the commonest 20,000 in the language, and miss many patterns. We are exploring using the web (suitably filtered) as the input corpus.

Acknowledgements

This work was supported by the UK EPSRC, under the WASPS project, grant GR/M54971. We would like to thank Prof. Tony Hartley from Leeds University for organising the experiments.

References

COBUILD, 1987. *The Collins COBUILD English Language Dictionary.* Edited by John McH. Sinclair *et al.* London.

William Gale, Kenneth Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(1–2):415–539.

Adam Kilgarriff and Michael Rundell. 2002. Lexical profiling software and its lexicographical applications - a case study. In *EURALEX 02*, Copenhagen.

Adam Kilgarriff and David Tugwell. 2001a. Waspbench: an MT lexicographer's workstation supporting state-of-the-art lexical disambiguation. In *Proc. MT Summit VIII*, pages 187–190, Santiago de Compostela, Spain, September.

Adam Kilgarriff and David Tugwell. 2001b. Word sketch: Extraction and display of significant collocations for lexicography. In *Proc. Collocations workshop, ACL 2001*, pages 32–38, Toulouse, France.

Adam Kilgarriff. 1998. The hard parts of lexicography. *International Journal of Lexicography*, 11(1):51–54.

Rob Koeling and Adam Kilgarriff. 2002. Evaluating the waspbench, a lexicography tool incorporating word sense disambiguation. In *Proceedings of ICON 2002*, Mumbai, India, December.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of ACL*, Montreal.

John M. Sinclair, editor. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing.* Collins, London.

David Yarowsky and Radu Florian. 2002. Evaluating sense disambiguation performance across diverse parameter spaces. *Journal of Natural Language Engineering*, page In press. Special Issue on Evaluating Word Sense Disambiguation Systems.

David Yarowsky. 1993. One sense per collocation. In *Proc. ARPA Human Language Technology Workshop*, Princeton.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of ACL 1995*, pages 189–196, Cambridge, MA.

Two Approaches to Aspect Assignment in an English-Polish Machine Translation System

Anna Kupść
Polish Academy of Sciences, Institute of Computer Science and
Carnegie Mellon University, Language Technologies Institute
aniak@cs.cmu.edu

Abstract

The paper presents two approaches to aspect assignment in a knowledge-based English-Polish machine translation (MT) system. The first method uses a set of heuristic rules based on interlingua (IR) representation provided by the system, whereas the other employs machine learning techniques. Both methods have similar performance and obtain high accuracy of over 88% on test data. The crucial difference, however, is the development effort: the machine learning technique is fully automatic, whereas heuristic rules are derived manually.

1 Introduction

The paper presents two methods to deal with aspect assignment in a prototype of a knowledge-based English-Polish machine translation (MT) system. Although there is no agreement among linguists as to its precise definition, e.g., Vendler (1967), Comrie (1976), Dowty (1986), aspect is a result of complex interplay of semantics, tense, mood and pragmatics and it strongly affects overall text understanding. In English, aspect is usually not explicitly indicated on a verb. On the other hand, in Polish it is overtly manifested and incorporated into verb morphology. This difference between the two languages makes English-Polish translation particularly difficult as it requires contextual and semantic analysis of the English input in order to derive aspect value for the Polish output.

The MT system presented in this paper takes advantage of a knowledge-based interlingua (IR) representation in order to assign aspect in Polish translation. We propose two approaches based on this representation. First, we provide a set of human-defined heuristic rules (similar to 'cues strategy' presented in Gawrońska (1993)), and second, we use machine learning techniques to learn aspect assignment rules. The former approach has been incorporated into the system, whereas the latter has been, so far, run separately as an experiment. The results obtained by both methods are quite similar. The crucial difference, however, is the effort put into their development: the machine learning approach is fully automatic and rules are derived from examples rather than hand-coded.

The organization of the paper is as follows: section 2 briefly presents the system architecture, sections 3 and 4 describe heuristic rules and the machine learning approach, respectively. Finally, section 5 contains conclusions.

2 System description

The English-Polish MT project presented in this paper is an extension of the existing multilingual KANTOO system (a reimplementation of the KANT system, cf. Mitamura et al. (1991), Mitamura and Nyberg (1992)) developed at Carnegie Mellon University. KANTOO is a knowledge-based, high-quality, domain-specific MT system (in the English-Polish MT project, the domain is restricted to printer manuals) and it uses Interlingua (IR) as a semantic representation, see Leavitt et al. (1994). The system takes as an input a text written in constrained English (controlled language), which limits vocabulary and grammar of sentences accepted by the system, cf. Kamprath et al. (1998). Example (1) presents a sample English input along with the IR representation and its Polish translation provided by the system.

```
(1)   The printer prints pages.

      (*A-PRINT
        (agent
          (*O-PRINTER
            (number singular)
            (reference
                   definite)))
        (argument-class
                  agent+theme)
        (mood declarative)
        (punctuation period)
        (tense present)
        (theme
          (*O-PAGE
            (number plural)
            (reference
              no-reference))))

      Drukarka drukuje strony.
```

IR illustrated in (1) is the input for the Polish generation module. The module consists of four components: a mapper, a unification grammar (a type of context-free grammar), a morphological generator and a post-processing module. Mapping rules transform the IR semantic representation into a syntactic structure corresponding to the Polish output. The structure is a functional structure or FS in the LFG (Lexical Functional Grammar) formalism, cf. Bresnan (1982). Generation grammar rules convert this FS into a list of lexical tokens (FS frames), which are then fed to the morphology module responsible for generating appropriate inflected forms. Finally, a set of post-processing rules is applied to produce the resulting surface form of translation by cleaning up spacing, adding capitalization, inserting punctuation, etc. In order to develop the current system, a small corpus of about 280 English sentences from a printer manual has been examined. This corpus served as a baseline to develop the two approaches to aspect assignment presented in the paper.

As mentioned above, aspect is incorporated into verb morphology in Polish. Polish verbs may have two aspect forms: imperfective, e.g., *drukuje* 'prints', or perfective, e.g., *wydrukuje* 'will print$_{3.sg}$ (out)'. Aspect is independent of tense or mood as it is also present on infinitives: *drukować* 'to print$_{imperf}$' and *wydrukować* 'to print$_{perf}$ (out)', or on gerunds: *drukowanie* 'printing$_{imperf}$' and *wydrukowanie* 'printing$_{perf}$ (out)'.

Since English verbs do not have morphological aspect, we consider lexical concepts, e.g., `*A-PRINT` in (1), ambiguous: they can be translated by either a perfective or an imperfective verb, see (2).

```
(2)   *A-PRINT = ([?verb] drukowac)
      drukowac =
      (*OR* ((morph verb-imperf)
              (root drukuje))
             ((morph verb-perf)
              (root wydrukuje))),
```

The role of aspect assignment rules is to specify which form to use in translation. The next two sections describe two methods which provide such rules based on IR specification.

3 Heuristic rules

Heuristic rules are specified in the mapper and they assign aspect according to attributes found in IR. The rules are ordered so that more general cases are considered first and if they do not hold, more specific rules are applied. Aspect assignment rules proposed in the system are presented below.

3.1 Declarative Mood

For finite verbs in declarative mood, aspect primarily depends on tense. First, all continuous forms, marked as `(progressive +)` in IR, are translated as imperfective. Next, forms of perfective tenses, i.e., `(perfective +)`, are translated as perfective. Then, verbs in simple past, `(tense past)`, or future simple tenses, `(tense future)`, are translated as perfective. Similar assignment rules have been proposed in Gawrońska (1993).

Additionally, we assume that certain types of subordinate conjunctions, e.g., ‘while’, ‘once’, ‘before’, etc., impose aspect requirements on a verb in the subordinate clause. The following assignments have been proposed:

● ‘while’: imperfective

(3) You can send an electronic fax **while** the printer makes copies.

Można wysłać elektroniczny faks,
can send$_{inf}$ electronic fax
podczas gdy drukarka robi
while printer makes$_{imperf}$
kopie.
copies

● ‘once’, ‘after’, ‘before’, ‘until’: perfective; additionally clauses introduced by the conjunction ‘until’ have to be negated in Polish

(4) Jobs also queue and wait **until** another job finishes.

Zadania także ustawiają się w kolejce
jobs also stand REFL in queue
i czekają, dopóki inne zadanie nie
and wait until another job not
skończy się.
finishes$_{perf}$ REFL

● ‘by’+gerund: imperfective; such clauses are translated into Polish by a contemporary adverbial participle derived only from imperfective verbs, see Saloni and Świdziński (1985)

(5) Close the document **by selecting** `Close` from the `File` menu.

Zamknij dokument wybierając
close document selecting$_{imperf}$
`Zamknij` z menu `Plik`.
`Close` from menu `File`

If none of the above cases hold, we assume that aspect of present tense verbs is imperfective. This assignment is valid also for gerunds as they are represented in IR as present tense verbs with an additional attribute `(nominal +)`.

3.2 Imperative Mood

After a brief analysis of Polish technical documentation, we decided to condition aspect in imperative mood on negation. Negated imperatives more often appear with imperfective forms (86.2%), whereas perfective aspect prevails with non-negated imperatives (83.5%).

Heuristic rules used in the system conform with the above statistics: we translate non-negated imperatives as perfective, (6), and negated imperatives as imperfective verbs, (7).

(6) Print a test page.

Wydrukuj stronę próbną.
print$_{perf}$ page test

(7) Do not move the lever after the scanner has begun sending the page.

Nie przesuwaj dźwigni, gdy skaner
not move$_{imperf}$ lever when scanner
zaczął wysyłanie strony.
started sending page

3.3 Infinitives

Infinitives have no mood or tense specified and we need separate rules to resolve aspect of these forms. In general, English infinitives appear as either complements of other verbs, e.g., modals, or they head infinitive clauses introduced by a conjunction such as ‘in order to’. We assume that in the former case, aspect of the infinitive depends on the governing verb while in the latter — on the subordinate conjunction.

For the conjunction ‘in order to’, we assume that it requires a perfective infinitive argument, (8).

(8) You must unhook the other device **in order to** connect the printer.

Trzeba wyłączyć inne urządzenie,
need unhook$_{perf}$ another device
aby podłączyć drukarkę.
in order to connect$_{perf}$ printer

Modal verbs are represented in IR by a set of semantic attributes such as `ability`, `possibility`, `tentativity`, `necessity`, `obligation`, see Leavitt et al. (1994). The following aspect assignment has been adopted in the system:

- ‘can’, `(ability +)` or `(possibility +)`: perfective;
- ‘cannot’, `(ability +) (negation +)`: imperfective;
- ‘cannot’, `(possibility +) (negation +)`: perfective;
- ‘could’, `(possibility +) (tentativity low)`: perfective;
- ‘may’, `(possibility +) (tentativity medium)`: perfective;
- ‘must’, `(obligation medium)`: perfective;
- ‘should’, `(expectation +)`: perfective.

3.4 Results

As mentioned above, aspect strongly depends on semantic and pragmatic context. Since such information is impoverished in KANTOO, the proposed rules cannot be perfect. In order to evaluate their performance, results obtained by the system have been compared with human translations of the initial (training) English corpus (280 sentences). The heuristic rules have been developed in order to accommodate data in the training corpus. Therefore, in order to obtain a more objective verification of the proposed rules, we additionally tested the system performance on a separate set of 24 (test) sentences taken from the same manual. The results obtained on training and test sets are summarized in Fig. 1.

result	train		test	
	#verbs	%	#verbs	%
correct	430	88.1%	53	88.3%
incorrect	58	11.9%	7	11.7%

Figure 1: Performance of heuristic rules

4 Machine Learning

The machine learning approach described in this section is also based on the IR representation provided by the MT system. In this experiment, we used the C4.5 software to build a decision tree. Training and test data have been derived from the same sentences the heuristic rules have been proposed for and evaluated on. We have run the experiment twice, using two different measures to build the decision tree: information gain and gain ratio. Performance of both algorithms has been evaluated on unpruned and pruned trees. Additionally, the optimal (pruned) trees have been transformed into rules and their accuracy has been measured as well. Details of the experiment and its results are presented below.

4.1 Data

Data used for training and testing were taken from the same set of sentences the heuristic rules have been applied to. All sentences have been analysed by the KANTOO analyser and the resulting IR served as an input for preparing the data. In particular, we have selected 12 attributes which had been crucial for development of the heuristic rules: `ability`, `expectation`, `marker`, `mood`, `necessity`, `negation`, `obligation`, `perfective`, `possibility`,`progressive`,`tense`,`tentativity`.

Most of these attributes are taken directly from IR, with an exception to `marker`, which has been introduced to indicate the type of subordinate conjunction, e.g., ‘while’, ‘unless’, ‘once’, etc. Note that not all attributes are specified in IR for every verb, e.g., infinitives do not have the `mood` attribute. We have slightly modified the mapper to make sure that all 12 attributes required for learning are present for every verb and have their values specified. Values of attributes missing in IR are either set to ‘`-`’ or `none`, depending on whether the attribute is binary or has more values. In addition, every verb in the data set has been labelled with the correct aspect value based on the human translation. The resulting 13-tuples served as training data for the decision tree. The target concept (aspect) has been represented by a binary attribute: `0` corresponds to imperfective, `1` to perfective aspect. The test data has the same format.

Due to changes in the mapper, the final number of examples used in the experiment was smaller than in the original system. The decision tree was trained on 417 and tested on 55 examples.

4.2 Decision Trees

As mentioned above, we employed two measures to build a decision tree: information gain, Quinlan (1986), and gain ratio, Quinlan (1986; Quinlan (1993). The main difference between the two techniques is in the size of the resulting tree: the former favours attributes with multiple values, which results in a wider (and usually bigger) tree. Indeed, the tree built according to gain ratio is smaller (31 nodes), Fig. 2, whereas the one based on information gain is slightly bigger (33 nodes), Fig. 3.

```
mood = none: 1
mood = declarative:
|   tense = past: 1
|   tense = future: 1
|   tense = none: 0
|   tense = present:
|   |   progressive = +: 0
|   |   progressive = -:
|   |   |   possibility = +: 1
|   |   |   possibility = -:
|   |   |   |   marker = to-inf: 1
|   |   |   |   marker = while: 0
|   |   |   |   marker = because: 0
|   |   |   |   marker = if: 0
|   |   |   |   marker = until: 1
|   |   |   |   marker = by_ing: 0
|   |   |   |   marker = in-order-to: 0
|   |   |   |   marker = after: 1
|   |   |   |   marker = when: 1
|   |   |   |   marker = unless: 1
|   |   |   |   marker = once: 1
|   |   |   |   marker = so_that: 1
|   |   |   |   marker = none:
|   |   |   |   |   ability = +: 1
|   |   |   |   |   ability = -:
|   |   |   |   |   |   obligation =
                                none: 0
|   |   |   |   |   |   obligation =
                              medium: 1
mood = imperative:
|   negation = +: 0
|   negation = -: 1
```

Figure 2: Decision tree based on gain ratio

The produced trees turned out to be optimal with respect to the learning algorithm (every node in the tree produced an improvement over the training data) and no nodes were pruned. Evaluation of the decision trees on the training and test data is summarized in Fig. 4.

```
mood = none: 1
mood = declarative:
|   marker = to-inf: 1
|   marker = while: 0
|   marker = because: 0
|   marker = if: 0
|   marker = until: 1
|   marker = by_ing: 0
|   marker = in-order-to: 0
|   marker = after: 1
|   marker = unless: 1
|   marker = once: 1
|   marker = so_that: 1
|   marker = none:
|   |   tense = past: 1
|   |   tense = future: 1
|   |   tense = none: 0
|   |   tense = present:
|   |   |   possibility = +:
|   |   |   |   progressive = +: 0
|   |   |   |   progressive = -: 1
|   |   |   possibility = -:
|   |   |   |   ability = +: 1
|   |   |   |   ability = -:
|   |   |   |   |   obligation =
                              none: 0
|   |   |   |   |   obligation =
                         medium: 1
|   marker = when:
|   |   progressive = +: 0
|   |   progressive = -: 1
mood = imperative:
|   negation = +: 0
|   negation = -: 1
```

Figure 3: Decision tree based on information gain

measure used in decision tree	error		
	train	estimate	test
gain ratio	9.8%	11.1%	10.9%
information gain	9.6%	10.8%	10.9%

Figure 4: Performance of decision trees

The error estimate presented in Fig. 4 indicates the predicted error rate on unseen examples (the so-called pessimistic estimate): the upper bound of the error based on the observed error on the training data for a given confidence level (set to 95% in the experiment). As shown in Fig. 4, the decision tree built according to gain ratio performed 0.2% worse on training data and it had 0.3% higher error estimate than the information gain tree. The gain ratio estimate overestimates the actual error on test (unseen) data by 0.2%, whereas the information gain estimate underestimates it by 0.1%. Hence, the results obtained by both classifiers are very similar and difference may be attributed to random noise. In order to eliminate this effect, they should be tested on a bigger sample, which was unavailable in the present experiment.

4.3 Automatically Learned Rules

The final part of the experiment consisted in converting the decision trees into rules and verify their performance. Initially, both trees were represented by the same number of rules (21) but after evaluation on the training data, one rule (Rule 18) has been removed from the gain ratio tree. The rules obtained from both decision trees are very similar but they appear in a different order and may have different accuracy, see Fig. 5 and Fig. 6. The rules are grouped according to their output class (i.e., aspect value), ordered with respect to accuracy within this class and applied in the obtained order. Examples to which none of the rules apply fall into the `default` class, computed individually for each tree. Performance of both sets of rules is identical: 9.4% errors on training and 10.9% errors on test data. Therefore, the learned rules score higher than the heuristic rules which have 11.9% errors on training and 11.7% errors on test data.

Note that the learned rules comprise the heuristic rules discussed in sec. 3. The only exception is Rule 5, which does not take into account `negation` and misclassifies complements of ‘cannot’ as perfective. Some of the heuristic rules do not have explicit counterparts among the learned rules. Heuristic rules referring to `perfective`, `tentativity`, `expectation` or the marker ‘before’ are not overtly present in the decision trees.

```
Rule 4:               Rule 5:
 marker = to-inf       ability = +
 -> class 1 [99.7%]    -> class 1 [70.6%]
Rule 13:              Rule 14:
 marker = after        marker = when
 -> class 1 [99.0%]    progressive = -
Rule 7:                -> class 1 [62.0%]
 obligation = medium  Rule 2:
 -> class 1 [98.3%]    progressive = +
Rule 11:               -> class 0 [99.8%]
 marker = until       Rule 12:
 -> class 1  [97.5%]   marker = by_ing
Rule 1:                -> class 0  [99.4%]
 mood = none          Rule 8:
 -> class 1 [96.3%]    marker = while
Rule 15:               -> class 0 [99.0%]
 marker = unless      Rule 20:
 -> class 1 [95.0%]    mood = imperative
Rule 16:               negation = +
 marker = once         -> class 0 [98.3%]
 -> class 1 [95.0%]   Rule 9:
Rule 21:               marker = because
 mood = imperative     -> class 0 [97.5%]
 negation = -         Rule 6:
 -> class 1 [93.2%]    ability = -
Rule 19:               marker = none
 tense = future        mood = declarative
 -> class 1 [78.2%]    obligation = none
Rule 18:               possibility = -
 tense = past          tense = present
 -> class 1 [74.6%]    -> class 0 [88.1%]
Rule 3:               Rule 10:
 possibility = +       marker = if
 progressive = -       -> class 0 [78.9%]
 -> class 1 [73.6%]   Default class: 1
```

Figure 5: Automatic rules for the gain ratio tree

Recall, however, that all these rules resolved aspect to perfective. In the machine learning approach, they are covered by the `default` rule. Finally, note that in the machine learning approach several new rules have been discovered: Rules 9, 10, 14 and 15 in Fig. 5 (11, 12, 16 and 17 in Fig. 6) do not correspond to any of the heuristic rules.

```
Rule 2:               Rule 5:
 marker = to-inf       ability = +
 -> class 1 [99.7%]    -> class 1 [70.6%]
Rule 8:               Rule 16:
 marker = none         marker = when
 tense = past          progressive = -
 -> class 1 [99.0%]    -> class 1 [62.0%]
Rule 15:              Rule 3:
 marker = after        progressive = +
 -> class 1 [99.0%]    -> class 0 [99.8%]
Rule 7:               Rule 14:
 obligation = medium   marker = by_ing
 -> class 1 [98.3%]    -> class 0 [99.4%]
Rule 13:              Rule 10:
 marker = until        marker = while
 -> class 1 [97.5%]    -> class 0 [99.0%]
Rule 1:               Rule 20:
 mood = none           mood = imperative
 -> class 1 [96.3%]    negation = +
Rule 17:               -> class 0 [98.3%]
 marker = unless      Rule 11:
 -> class 1 [95.0%]    marker = because
Rule 18:               -> class 0 [97.5%]
 marker = once        Rule 6:
 -> class 1 [95.0%]    ability = -
Rule 21:               marker = none
 mood = imperative     mood = declarative
 negation = -          obligation = none
 -> class 1 [93.2%]    possibility = -
Rule 9:                tense = present
 tense = future        -> class 0 [88.1%]
 -> class 1 [78.2%]   Rule 12:
Rule 4:                marker = if
 possibility = +       -> class 0 [78.9%]
 progressive = -      Default class: 1
 -> class 1 [73.6%]
```

Figure 6: Automatic rules for the information gain tree

5 Conclusions

In this paper, we presented two approaches to aspect assignment in a knowledge-based English-Polish MT system: heuristic rules and machine learning. As for approaches which do not rely on semantics or pragmatics, accuracy of both methods is very high: heuristic rules achieve 88.3% and automatically learned rules 89.1% accuracy on test data. Although the final results turn out to be very similar, the crucial difference between the two methods is the development effort: the machine learning technique acquires rules automatically, while heuristic rules are hand-coded. Another advantage of the machine learning approach is that it allows for more concise encoding of the heuristic rules and discovering new rules.

It has to be noted that the success of the machine learning approach strongly relies on the choice of attributes used for learning. The heuristic rules and the decision trees employ the same attributes. Therefore, human knowledge is necessary to limit the search space in the automatic approach. Another factor which contributed to the high system performance is the restricted domain of translation and use of controlled language. Although some heuristics are quite general (e.g., the rules compatible with those independently proposed in Gawrońska (1993)), the system probably will not be fully scalable to an open-domain unrestricted natural language text. Providing reliable heuristics in a general purpose MT system will be much more difficult than for a domain-specific MT system. On the other hand, having set the learning attributes (or corresponding surface / syntactic patterns), machine learning methods can be successfully applied to automatically acquire rules from annotated data.

Acknowledgments

I wish to thank Krzysztof Czuba, Kathryn L. Baker and two anonymous reviewers of the EACL EAMT Workshop for their comments and suggestions to improve this paper.

References

Joan Bresnan, editor. 1982. *The Mental Representation of Grammatical Relations*. MIT Press Series on Cognitive Theory and Mental Representation. The MIT Press, Cambridge, MA.

Bernard Comrie. 1976. *Aspect. An introduction to the study of verbal aspect and related problems*. Cambridge University Press, Cambridge.

David Dowty. 1986. The effects of aspectual class on the temporal structure of discourse: Semantics or pragmatics? *Linguistics and Philosophy*, 9:37–61.

Barbara Gawrońska. 1993. *An MT Oriented Model of Aspect and Article Semantics*. Lund University Press.

Christine Kamprath, Eric Adolphson, Teruko Mitamura, and Eric Nyberg. 1998. Controlled language for multilingual document production: Experience with Caterpillar technical English. In *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW '98)*. Available from http://www.lti.cs.cmu.edu/Research/Kant/claw98ck.pdf.

John R. Leavitt, Deryle W. Lonsdale, and Alexander M. Franz. 1994. A reasoned interlingua for knowledge-based machine translation. In *Proceedings of CSCSI-94*. Available from: http://www.lti.cs.cmu.edu/Research/Kant/.

Teruko Mitamura and Eric Nyberg. 1992. The KANT system: Fast, accurate, high-quality translation in practical domains. In *Proceedings of COLING-92*.

Teruko Mitamura, Eric Nyberg, and Jaimie Carbonell. 1991. An efficient interlingua translation system for multi-lingual document production. In *Proceedings of the Third Machine Translation Summit*.

J. Ross Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Zygmunt Saloni and Marek Świdziński. 1985. *Składnia Współczesnego Języka Polskiego*. Państwowe Wydawnictwo Naukowe, Warszawa, 2nd edition.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY.

Multi-language Machine Translation through Interactive Document Normalization

Aurélien Max
Groupe d'Etude pour la Traduction Automatique (GETA)
Xerox Research Centre Europe (XRCE)
Grenoble, France
aurelien.max@imag.fr

Abstract

Document normalization is an interactive process that transforms raw legacy documents into semantically well-formed and linguistically controlled documents with the same communicative intention content. A paradigm for content analysis has been implemented to select candidate semantic representations of the communicative content of an input document. This implementation reuses the formal content specification of a multilingual controlled authoring system. As a consequence, a candidate semantic representation can not only be associated with a text in the language of the input document, but also in all the languages supported by the system. This paper presents how multilingual versions of an input legacy document can be obtained interactively with a proposed implementation, and discusses the advantages and limitations of this kind of *normalizing translation*.

1 Introduction

Translating unrestricted text by machine is a problem that has been involving a lot of research for the past decades, but is still far from solved (Cole et al., 1996). This task arises so many problems in computational linguistics, most of them only partially solved, that a lot of research is still to be carried out before one can ask a personal computer to translate accurately an arbitrary piece of text from one language to another. The performance bottleneck due to the lack of linguistic and knowledge resources has led the builders of practical translation systems to constrain the input to controlled languages, and/or to have recourse to human expertise on the source language and on the discourse domain (e.g. (Boitet and Blanchon, 1996; Baker et al., 1994)). Unsurprisingly, the most successful systems to date operate on text in very limited domains, exemplified by the weather forecast translation from English to French of the METEO system.

There exist many situations where documents belonging to a constrained domain have to be translated in several languages, as is the case of official documents in multilingual communities or product descriptions for international companies. In these situations one have at least the following expectations:

- **high-quality translation**, which implies that it be accurate and not necessarily literal
- outputs in possibly **many target languages**
- **consistency** across documents of the same class (e.g. drug leaflets, experiment reports), so that concepts are always expressed in the same unambiguous manner and the texts produced can be regarded as **gold standards** for the meaning they convey

Differents methods that do not impose constraints on the input text have been proposed to achieve high-quality translation. Interaction with a user can be used to disambiguate the input text, and could be prefered to to post-editing as this has to be done only once for all languages, thus reducing the time and efforts needed. Interlingual representations (Hutchins and Somers, 1992) are well-adapted to support the production of the target text in several languages, and they can also be effectively used to check the semantic coherence and well-formedness of a document. Reusing previous translations, as proposed in the different flavours of Example-based Machine Translation (Somers, 1999), is an interesting alternative to purely rule-based approaches and allows the selection of non-literal high-quality translation candidates.

This paper starts with a short presentation of an authoring system that allows the creation of multingual documents with all the above properties. *Document normalization*, which is described next, stemmed from the question of whether such an authoring system could be used in a reversed mode to analyze existing documents from the class of documents supported by it. After providing a motivating example, we will briefly introduce *fuzzy inverted generation*, a paradigm we proposed to normalize documents reusing the formalism of the abovementioned authoring system, and describe a document normalization system. We will then attempt to define how *normalizing translation* can be achieved through document normalization, and we will discuss the advantages and limitations of such an approach.

2 Controlled Document Authoring

Controlled Document Authoring is an active field of research comprising approaches such as the *What You See Is What You Meant* (WYSIWYM) paradigm (Power and Scott, 1998) and *Multilingual Document Authoring* (MDA) (Dymetman et al., 2000). The systems allow authors to specify document content representations interactively in their own language, and then produce versions in several languages using parallel resources.

In MDA, a system developed at XRCE, the author of a document has to select valid semantic choices in active fields interspersed with the evolving text of the document in her language until the document is complete (see figure 1). The system can at any time produce current versions of the documents from the content representation in all the languages it supports. The documents thus obtained are of high-quality, and are not necessarily literal translations but rather adaptations to a given language.[1] In fact, the linguistic structures of two documents can be completely different in two different languages, and communicative intentions can be conveyed in quite different ways. Moreover, since the generator of an MDA system is deterministic, the texts produced will be consistent across documents.

The specification of well-formed document content representations in MDA is recursively described in a grammar formalism that is a variant of Definite Clause Grammars (Pereira and Warren, 1980). Text strings can appear in right-hand sides of rules, which allows text realizations to be associated to content representations, and thus provides a close coupling between semantic modelling and generation. Figure 2 shows an abstract typed tree in the MDA formalism and its realizations as English and French sentences.[2] Non-terminals are typed semantic elements whose type appears after the two colons. Dependencies can be enforced through the use of shared variables between semantic elements. The granularity of text fragments in rules is not necessarily a fine-grained predicate-argument structure of sentences commonly used in NLG, so this is an intermediate level between full NLG and templates (Reiter, 1995). This approach proved to be adequate for classes of documents where the productivity of certain choices could be rendered as entire text spans, as is the case for example of warning sections in drug leaflets (Brun et al., 2000).

3 Document normalization

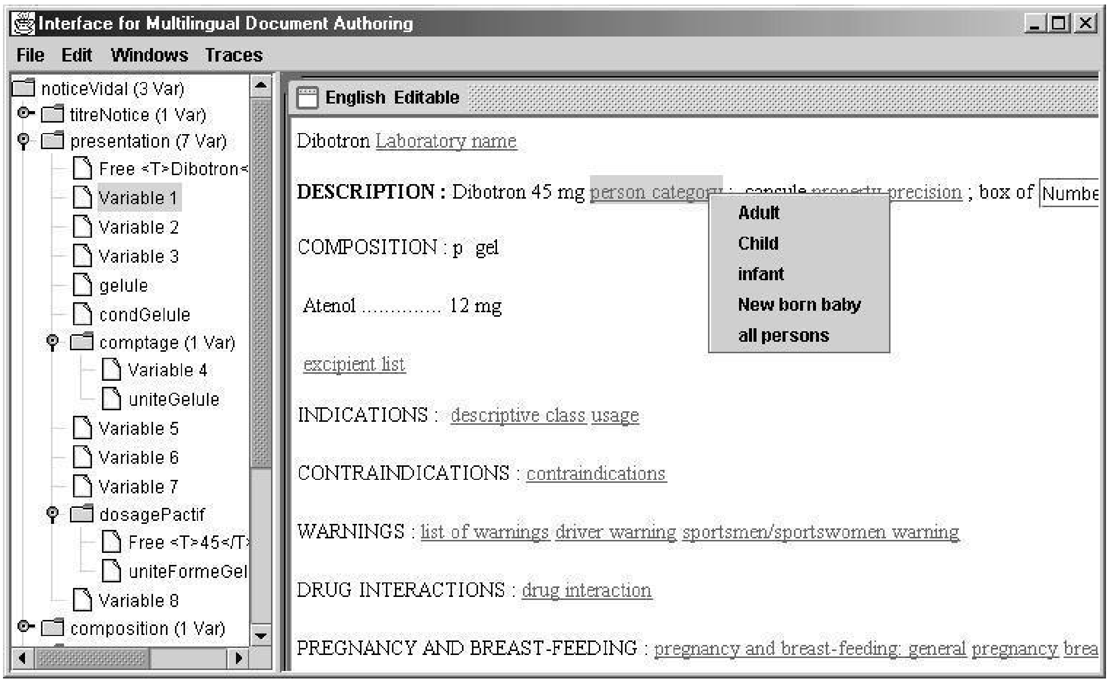
3.1 A Motivating example

The pharmaceutical domain produces yearly publications which are compendiums of documents initially produced by pharmaceutical companies which are presented in a consistent way (e.g. (ABPI, 1996; OVP Editions du VIDAL, 1998)). Several kinds of variations were observed in a corpus study we conducted on a corpus of 50 patient pharmaceutical leaflets for pain relievers from different drug vendors (Max, 2002). First, the structures of the leaflets could vary considerably, as well as the locations where certain communicative goals were expressed. (Paiva, 2000) showed the presence of significant stylistic variation in a corpus of 342 patient leaflets. Our study also revealed that similar communicative intentions could be expressed in a variety of ways conveying more or less subtle semantic distinctions. Seeing the content of such documents as goal-driven communication, a given utterance can be seen as an attempt to satisfy some communicative goal on the part of the writer of the document. We argue that for documents of the importance of pharmaceutical leaflets consistency of expression and of information presentation can be beneficial to the reader by allowing a clear and unambiguous understanding of the communicative goals contained in different documents. It can indeed sometimes be confusing for a reader to find various ways to express the same communicative intentions, as in the following examples:

- *This product should not be taken for more than*

[1] Different parts of the document can thus be easily localized: for example, disclaimers and contact information can be adapted to the targeted community.

[2] This example and its specification are inspired from the Nespole! project, a speech-to-speech translation project.



Interface for Multilingual Document Authoring

File Edit Windows Traces

noticeVidal (3 Var)
titreNotice (1 Var)
presentation (7 Var)
Free <T>Dibotron<
Variable 1
Variable 2
Variable 3
gelule
condGelule
comptage (1 Var)
Variable 4
uniteGelule
Variable 5
Variable 6
Variable 7
dosagePactif
Free <T>45</T>
uniteFormeGel
Variable 8
composition (1 Var)

English Editable

Dibotron Laboratory name

DESCRIPTION : Dibotron 45 mg person categor ; capsule property precision ; box of Numbe

Adult
Child
infant
New born baby
all persons

COMPOSITION : p gel

Atenol .... 12 mg

excipient list

INDICATIONS : descriptive class usage

CONTRAINDICATIONS : contraindications

WARNINGS : list of warnings driver warning sportsmen/sportswomen warning

DRUG INTERACTIONS : drug interaction

PREGNANCY AND BREAST-FEEDING : pregnancy and breast-feeding: general pregnancy brea

Figure 1: View of the MDA system during the authoring of a patient information leaflet

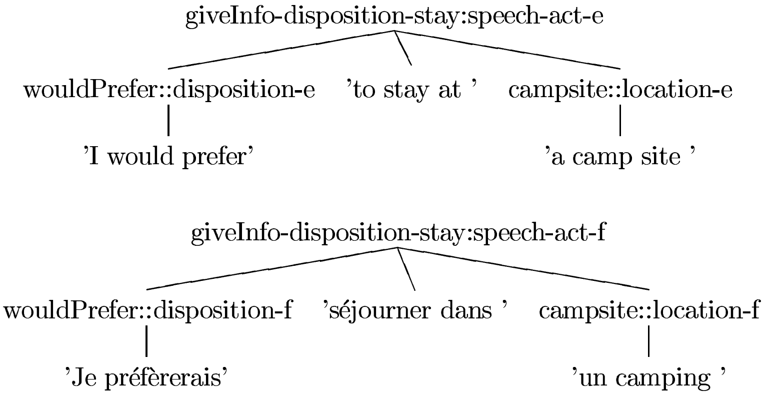


Figure 2: Abstract typed trees in English and French for the sentence *I would prefer to stay at a camp site*

14 days without first consulting a health professional.

- *If pain persists after 14 days, consult your doctor before taking any more of this product.*
- *If symptoms persist for 2 weeks, stop using this product and see a physician.*

Document normalization can be achieved by analyzing a legacy document into a semantically possible content representation, and producing a normalized version from that content representation. This normalized version expresses *predefined content*, which is conveyed in the input document, in a structurally and linguistically controlled way. Predefined content reveals *communicative goals*, which should typically be described by an expert of the discourse domain. Control on the production of text from some content representation allows to produce messages that can be seen as some sort of 'gold standard' for the communicative goal that are conveyed and that can be augmented to be made *self-explaining* (Boitet, 1996), and to obtain consistent document structures as well as to impose terminological and stylistic guidelines.

3.2 Fuzzy inverted generation

For the purpose of document normalization we would like to match texts that do not carry significant communicative differences in a given class of documents but may be of quite different surface forms. Therefore, we proposed to concentrate more on what counts as a well-formed document semantic representation rather than on surface properties of text, as the space of possible content representations is vastly more restricted than the space of possible texts.

Bridging the gap between deep content and surface text can be done by using the textual predictions made by the generator of an MDA system from well-formed content representations to match an input document. Indeed, an MDA system can be used as a formal device for enumerating well-formed document representations in a constrained domain and associating textual representations with them. If we can compute a relevant measure of semantic similarity between the text produced for any document content representation and the text of a legacy document, we could possibly consider the representations with the best similarity score as those best corresponding to the legacy document under analysis. Since this kind of analysis uses predictions made by a natural language generator, we named it *inverted generation* (Max and Dymetman, 2002) (see fig. 3). We also qualified it *fuzzy*, because as a generator will seriously undergenerate with respect to all the texts that could be normalized to the same communicative intention, the matching procedure has to be performed at a more abstract level than on raw text to evaluate commonality of communicative content. Considering the types of documents that could be analyzed using this paradigm, it seemed relevant to expand the generative power of the system, so that different texts could be associated with the same content representations to increase the robustness of the analysis. Although this non-determinism proves beneficial for inverted generation, we implemented it in such a way that the generation process would still be done deterministically.

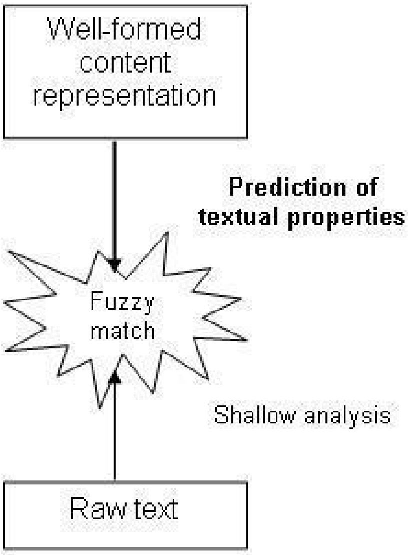

Well-formed content representation
Prediction of textual properties
Fuzzy match
Shallow analysis
Raw text

Figure 3: Deep content analysis through fuzzy inverted generation

To normalize an input document, we would like to find the *virtual document*[3] that is most similar to the input document in terms of the communicative content it conveys. The space of virtual documents for a given class of documents being potentially huge, we proposed an admissible heuristic search procedure (Nilsson, 1998), so that the candidate structures are returned in an order of decreasing similarity with the input text. The evaluation function it uses is an optimistic measure of similarity that corresponds to a weighted intersection between the *lexical profile* of the input

[3] We call *virtual document* a document that can be predicted by the authoring system but does not exist *a priori*.

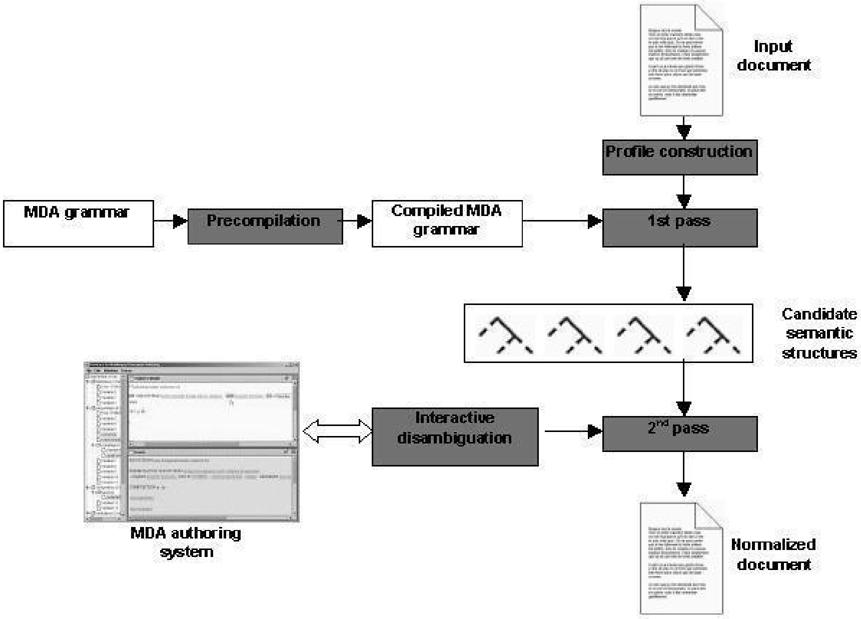


Figure 4: Architecture of the document normalization system

document and that computed for a partial content representation.[4] The lexical profile for a text fragment is defined as a vector of informative synsets[5] associated with their number of occurrences, and the lexical profile for an MDA semantic type gives the maximum number of occurrences of any given synset that could be attained by performing any derivation from that type.

3.3 Document normalization system

Figure 4 shows an overview of the document normalization system that we have started to develop. An MDA grammar is first compiled off-line to associate profiles with all its semantic types by percolating profiles in the grammar from the terminals up to the root type. This compiled version of the grammar is used in conjunction with the profile computed for the input document in a first pass analysis. The aim of this first pass analysis implementing fuzzy inverted generation is to isolate a limited set of candidate content representations. A second pass analysis is then applied on those candidates, which are now actual texts associated with their content representation, using more fine-grained linguistic analysis, in conjunction with interactive disambiguation when needed.[6].

4 Normalizing translation

Using the resources of a multilingual authoring system to analyze a legacy document offers a natural possibility: once the semantic content representation is obtained through document normalization, the generative capability of the authoring system can be reused to produce the documents corresponding to that representation in all the supported languages (see figure 5). Normalizing translation uses the same resources for both analysis and generation, and shares some properties with a pivot approach. A significant difference with previous approaches to translation using *reversible grammars* is that fuzzy matching is used. As this approach to machine translation relies on the matching with existing texts (those that can be produced by the generator of the authoring system), it shares some properties with Example-based Machine Translation (Somers, 1999), with the specificity that matched text fragments correspond first to semantic types in the MDA formalism and then eventually to their appropriate trans-

[4] More details on how fuzzy inverted generation can be implemented in MDA can be found in (Max, 2002).

[5] Synsets from WordNet, or ideally from a specialized thesaurus, have been prefered to lemma in order to account for lexico-semantic variation (Gonzalo et al., 1998).

[6] Typically, interactive disambiguation will allow an expert to prefer one of several ambiguous candidates on the basis of the legacy document.

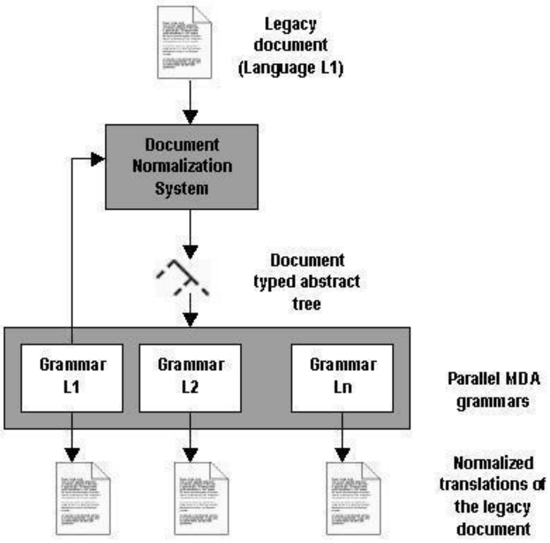

Legacy document (Language L1)
Document Normalization System
Document typed abstract tree
Grammar L1
Grammar L2
Grammar Ln
Parallel MDA grammars
Normalized translations of the legacy document

Figure 5: Normalizing translation using MDA grammars

lations in other languages. Under the assumption that the authoring system produces high-quality documents in all the languages it supports, then evaluating normalizing translation can be limited to evaluating the performance of document normalization. Forthcoming publications will attempt to address this issue.

A simple example of normalizing translation is given on figure 6. The discourse domain is assumed to be that of travel information, where some semantic distinctions are considered uninformative. In this case the English speaker wishes to mention his *family's first choice*, but this is lost in the translation. The utterance is best matched with wouldPrefer-3, a possible English realization for the semantic type disposition, in the context of giveInfo-disposition-stay. It is then normalized to the deterministic choice of the English generator, wouldPrefer. The structure to which it belongs, which is of type speech-act, can then be rendered in English as *I would prefer to stay at a camp site.* Using parallel MDA grammars for French and Spanish allows to obtain the corresponding abstract trees from which the French and Spanish versions of the normalized sentence can be obtained.

5 Discussion

The proposed approach to translation has important limitations: first, only documents in constrained domains which can be modeled with the MDA formalism can be dealt with. This excludes arbitrary pieces of text, and requires an initial development of the grammatical resources and its transposition to as many languages as the system should support. However, this would allow to reuse the document modeling for authoring new documents from scratch, which could modify in a beneficial way the documentation practices of technical writers. As opposed to 'traditional' machine translation, normalizing translation can only translate those elements of a text that fit in well-formed document content representations. Consequently, elements that are not modeled in the grammar used for analysis will not appear in the normalized version of the document and its translations, which makes normalizing translation performing a kind of content selection. Another delicate aspect of this approach is that if the normalization goes wrong, even though an expert could control and validate the whole process[7], then the multilingual versions of the resulting document will not be accurate translations of the input document.

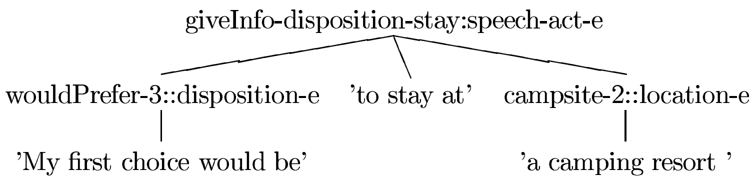
Despite the limitations given above, we think that this approach proposes enough advantages to be a viable solution for some well-defined contexts. First and foremost, if normalization goes well, so will translation, provided the parallel grammars of the authoring system are correct. The fuzzy inverted generation we have proposed has the inherent property of only producing candidates that are semantically well-formed and coherent, thus providing a means to the expert to correct or to reject an ill-formed legacy document. Validated documents are richer than usual textual documents since they are associated with their semantic description, which can for example be used to index the documents in a knowledge base for subsequent retrieval.

On the architectural side, the same resource is used for both analysis and generation, thus reducing considerably development time. Moreover, the fuzzy approach and the non-determinism of the inverted generation makes it possible to match a large range of inputs that could be more difficult to recognize using more traditional approaches to content analysis, such as syntactic parsing followed by semantic composition (Allen, 1995). Provided the necessary resources are available, notably a lemmatizer, a lexico-semantic database such as WordNet, and a human expert fluent in the appropriate language, any grammar of the authoring system could be used for analysis, therefore pos-

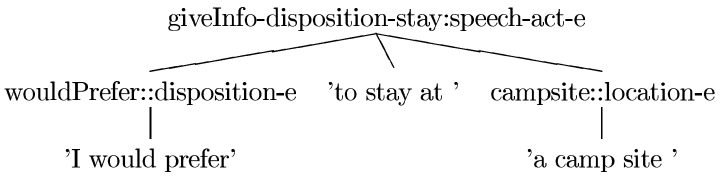
[7] The normalized document in the original language can be used by the expert to validate the normalization process, similarly to *feedback texts* in WYSIWYM.

Input text: *Staying at a camping resort is always my family's first choice.*

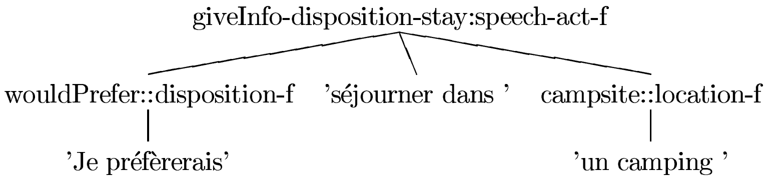
Best matching English abstract tree:


giveInfo-disposition-stay:speech-act-e
wouldPrefer-3::disposition-e
'to stay at'
campsite-2::location-e
'My first choice would be'
'a camping resort '

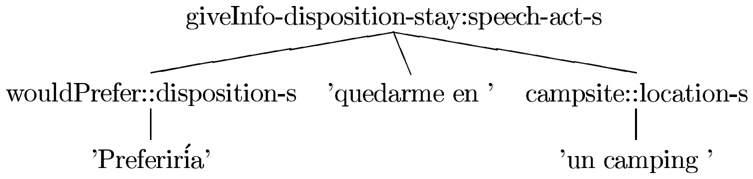
Corresponding normalized English abstract tree:


giveInfo-disposition-stay:speech-act-e
wouldPrefer::disposition-e
'to stay at '
campsite::location-e
'I would prefer'
'a camp site '

Corresponding French abstract tree:



Corresponding Spanish abstract tree:


giveInfo-disposition-stay:speech-act-s
wouldPrefer::disposition-s
'quedarme en '
campsite::location-s
'Preferiría'
'un camping '

Input text normalized translations:

- English: *I would prefer to stay at a camp site.*
- French: *Je préfèrerais séjourner dans un camping.*
- Spanish: *Preferiría quedarme en un camping.*

Figure 6: Example of normalizing translation for the English sentence *Staying at a camping resort is always my family's first choice* in the context of travel information

sibly allowing an N-to-N normalizing translation architecture. Finally, if the system has some supervised learning ability, for example by augmenting its generative power with examples validated by the expert, then it could be expected to perform better as more normalizations are done, as is the case with translation memories.

Acknowledgements The author wishes to thank Marc Dymetman and Christian Boitet for their supervision of his PhD work. This work is funded by a grant from ANRT.

References

ABPI. 1996. *ABPI Compendium of Patient Information Leaflets.* Datapharm Publications.

James Allen. 1995. *Natural Language Understanding.* Benjamin/Cummings Publishing, Redwood City, 2nd edition.

Kathryn L. Baker, Alexander M. Franz, Pamela W. Jordan, Teruko Mitamura, and Eric H. Nyberg. 1994. Coping with Ambiguity in a Large-Scale Machine Translation System. In *Proceedings of COLING-94, Kyoto, Japan.*

Christian Boitet and Hervé Blanchon. 1996. Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup. *Machine Translation*, 9:99–132.

Christian Boitet. 1996. Dialogue-based machine translation for monolinguals and future self-explaining documents. In *Proceedings of MIDDIM-96, Le Col de Porte, France.*

Caroline Brun, Marc Dymetman, and Veronika Lux. 2000. Document Structure and Multilingual Authoring. In *Proceedings of INLG 2000, Mitzpe Ramon, Israel.*

Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors. 1996. *Survey of the State of the Art in Human Language Technology.* Cambridge University Press.

Marc Dymetman, Veronika Lux, and Aarne Ranta. 2000. XML and Multilingual Document Authoring: Convergent Trends. In *Proceedings of COLING 2000, Saarbrucken, Germany.*

Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarrán. 1998. Indexing with WordNet Synsets Can Improve Text Retrieval. In *Proceedings of the COLING/ACL Workshop on the Usage of WordNet in Natural Language Processing Systems.*

W.J. Hutchins and Harold Somers. 1992. *An Introduction to Machine Translation.* Academic Press, London.

Aurélien Max and Marc Dymetman. 2002. Document Content Analysis through Inverted Generation. In *Proceedings of the workshop on Using (and Acquiring) Linguistic (and World) Knowledge for Information Access of the AAAI Spring Symposium Series, Stanford University, USA.*

Aurélien Max. 2002. Normalisation de Documents par Analyse du Contenu à l'Aide d'un Modèle Sémantique et d'un Générateur. In *Proceedings of TALN-RECITAL 2002, Nancy, France.*

Nils J. Nilsson. 1998. *Artificial Intelligence: a New Synthesis.* Morgan Kaufmann, San Francisco.

OVP Editions du VIDAL, editor. 1998. *Le VIDAL de la famille.* Hachette, Paris.

Daniel S. Paiva. 2000. Investing Style in a Corpus of Pharmaceutical Leaflets: Result of a Factor Analysis. In *Proceedings of the ACL Student Research Workshop, Hong Kong.*

Fernando Pereira and David Warren. 1980. Definite Clauses for Language Analysis. *Artificial Intelligence*, 13.

Richard Power and Donia Scott. 1998. Multilingual Authoring using Feedback Texts. In *Proceedings of COLING/ACL-98, Montreal, Canada.*

Ehud Reiter. 1995. NLG Vs Templates. In *Proceedings of ENLGW-95, Leiden, The Netherlands.*

Harold Somers. 1999. Review Article: Example-based Machine Translation. *Machine Translation*, 14:113–157.

Parallel Corpora Segmentation Using Anchor Words*

Francisco Nevado and **Francisco Casacuberta** and **Enrique Vidal**
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
{fnevado,fcn,evidal}@iti.upv.es

Abstract

A new technique for monotone segmentation of parallel corpora is introduced. This segmentation is based on a set of anchor words which are defined manually. The parallel segments are computed using a dynamic programming algorithm. To assess this technique, finite-state transducers are inferred from both non-segmented and segmented corpora. Experiments have been carried out with Spanish-English and Italian-English translation tasks. This technique has proven useful in improving the results with respect to those obtained with unsegmented corpora.

1 Introduction

In this paper, we present a new technique for improving machine translation systems. This is a heuristic approach for parallel corpora segmentation using anchor words and a dynamic programming algorithm.

In a parallel corpus, the anchor words are specific words that are defined for the two languages of the corpus and that are strongly related.

The goal of parallel corpus segmentation is to segment the source sentence and the target sentence in such a way that the correspondence between segments is monotone and one-to-one.

Using this segmentation, we attempted to improve the word alignments obtained with statistical techniques (Brown et al., 1993; Brown et al., 1990). These models depend on the length of the source and target sentences. The models are better estimated with shorter segments and, consequently, better word alignments are obtained.

The basic scheme of the proposed parallel segmentation is the following:

a) The source and the target sentences are initially segmented in the positions of the anchor words.

b) As the number of source and target segments can be different, a dynamic programming algorithm is applied to find the optimal correspondences between segments.

In section 2, we will show how to segment a bilingual corpus describing the segmentation of a pair of sentences using anchor words. We will then describe the experiments carried out to test this new technique and the obtained results.

2 Segmentation of a parallel corpus

Parallel segmentation is considered from a statistical point of view. Segmentation of a parallel corpus is carried out by segmenting every pair of sentences in this corpus.

2.1 Statistical machine translation

We use a notation which is similar to the one proposed in (Brown et al., 1993), where **f** is a source sentence and **e** is a target sentence.

*This work has been partially supported by TransType 2 (IST-2001-32091) and SISHITRA (TIC2000-1599-C02).

In order to translate from the source language to the target language in a statistical framework (Brown et al., 1993), we look for the probability of obtaining a sentence **e** from a sentence **f**, that is, $\Pr(\mathbf{e} \mid \mathbf{f})$. Applying Bayes rule, we have:

$$\Pr(\mathbf{e} \mid \mathbf{f}) = \frac{\Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})}{\Pr(\mathbf{f})}. \qquad (1)$$

Since we are searching for the target sentence with the best probability of being generated from the source sentence, by maximizing the preceding expression, we have:

$$\begin{aligned} \hat{\mathbf{e}} &= \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e} \mid \mathbf{f}) \\ &= \operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e}), \qquad (2) \end{aligned}$$

where $\Pr(\mathbf{e})$ corresponds to the probability of the target language model and $\Pr(\mathbf{f} \mid \mathbf{e})$ is known as the probability of the translation model. This model transforms a sentence in the target language into a sentence in the source language.

2.2 Segmentation of a pair of sentences

We obtain the segmentation as a byproduct of the translation process of a sentence. To start with, a trivial monolingual anchor-point-based **initial segmentation** is assumed on the sentence **f**. A different trivial monolingual anchor-point-based initial segmentation is also assumed on the sentence **e**. Having defined a set of anchor words for the source language and another set of anchor words for the target sentence, the first initial segment for sentence **f** is composed of the sequence of words from the beginning of the sentence until the first anchor word of **f**. The rest of the initial segments are composed of the sequences of words from the first word following the last segment until the next anchor word. The last segment of the sentence may end with the end of the sentence instead of an anchor word. The initial segments of **e** are computed in the same way, but taking into account the anchor words for the target language. Let us suppose that there are a initial segments for sentence **f** and there are b initial segments for sentence **e**. This initial segmentation is represented as:

$$\begin{aligned} \mathbf{e} &= \bar{e}_1 \bar{e}_2 \cdots \bar{e}_a &= \bar{e}_1^a \\ \mathbf{f} &= \bar{f}_1 \bar{f}_2 \cdots \bar{f}_b &= \bar{f}_1^b \end{aligned}$$

where $\bar{e}_c$ is a segment of consecutive words of **e** and $\bar{f}_d$ is a segment of consecutive words of **f**. See Figure 2 for an example of this kind of initial segmentation. $\bar{e}_{c_1}^{c_2}$ is the sequence of words constituted by the concatenation of the segments $\bar{e}_{c_1} \bar{e}_{c_1+1} \cdots \bar{e}_{c_2}$, and $\bar{f}_{d_1}^{d_2}$ is the sequence of words constituted by the concatenation of the segments $\bar{f}_{d_1} \bar{f}_{d_1+1} \cdots \bar{f}_{d_2}$.

Processing each initial segment as an atomic block, we can rewrite expression (2) with this notation:

$$\hat{\mathbf{e}} = \operatorname*{argmax}_{\bar{e}_1^a} \Pr(\bar{e}_1^a) \Pr(\bar{f}_1^b \mid \bar{e}_1^a). \qquad (3)$$

A parallel segmentation **s** is an ordered set of pairs of sequences of words, where every one of these pairs has a sequence of words from the source sentence and a sequence of words from the target sentence composed by one or more consecutive initial segments of the source sentence or the target sentence, respectively[1].

Given an initial segmentation $(\bar{e}_1^a, \bar{f}_1^b)$, we can represent a parallel segmentation as:

$$\mathbf{s} \equiv \Big([\bar{e}_1^{c_1}, \bar{f}_1^{d_1}], [\bar{e}_{c_1+1}^{c_2}, \bar{f}_{d_1+1}^{d_2}], \\ \ldots, [\bar{e}_{c_{|\mathbf{s}|-1}+1}^{c_{|\mathbf{s}|}}, \bar{f}_{d_{|\mathbf{s}|-1}+1}^{d_{|\mathbf{s}|}}] \Big),$$

where $|\mathbf{s}|$ is the number of segments for the parallel segmentation **s**. Clearly we have $c_{|\mathbf{s}|} = a$ and $d_{|\mathbf{s}|} = b$. Therefore, in a segmentation, any segment in the input sentence cannot be left without a corresponding segment of the output sentence, and vice versa. Another restriction is that there cannot be inversions in the order of the initial segments; that is, if $[\bar{e}_{c_1}^{c_2}, \bar{f}_{d_1}^{d_2}]$ is a pair of segments of a segmentation **s**, then $\forall [\bar{e}_{c_3}^{c_4}, \bar{f}_{d_3}^{d_4}] \in \mathbf{s}$:

$$\begin{aligned} &\text{if } c_2 < c_3 \Longrightarrow d_2 < d_3 \\ &\text{if } c_4 < c_1 \Longrightarrow d_4 < d_1 \end{aligned}$$

An example of this kind of segmentation is shown in section 2.3.

The set of possible parallel segmentations for an initial segmentation based on anchor words

[1] Note the difference with an initial segmentation. A segmentation may have joined several consecutive segments of the initial segments, but it has the same number of final segments for the source and target sentences.

$(\bar{e}_1^a, \bar{f}_1^b)$ is denoted by $\mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)$. Now, we can write the probability for the translation model, $\Pr(\bar{f}_1^b \mid \bar{e}_1^a)$:

$$\Pr(\bar{f}_1^b \mid \bar{e}_1^a) = \sum_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} \Pr(\bar{f}_1^b, \mathbf{s} \mid \bar{e}_1^a) \quad (4)$$

where $\Pr(\bar{f}_1^b, \mathbf{s} \mid \bar{e}_1^a)$ allows for the interpretation of a segmentation as a generative model. We can say that the segments in the source sentence are generated from the corresponding segments of the target sentence.

Given a sentence $\bar{e}_1^a$, we define the probability for a sentence $\bar{f}_1^b$ and a segmentation $\mathbf{s}$ as:

$$\Pr(\bar{f}_1^b, \mathbf{s} \mid \bar{e}_1^a) = \Pr(\mathbf{s} \mid \bar{e}_1^a) \prod_{q=1}^{|\mathbf{S}|} \Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}), \quad (5)$$

where $\Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q})$ is again the probability of the translation model for a subsequence of the sentence **f** and a subsequence of the sentence **e**.

We do not want to consider the translation model as a recursive model, so we will approximate the probability $\Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q})$ using Model 1 proposed in (Brown et al., 1993). In an intuitive manner, Model 1 computes the probability of a sequence of words to be translated by other sequence of words, without taking into account the word order. Therefore, it can allow translation inversions inside the sequences of words. The translation probability of a sequence of words $\bar{e}_{c_1}^{c_2}$ into a sequence of words $\bar{f}_{d_1}^{d_2}$ using Model 1 is computed by:

$$\Pr(\bar{f}_{d_1}^{d_2} \mid \bar{e}_{c_1}^{c_2}) = M_1(\bar{f}_{d_1}^{d_2} \mid \bar{e}_{c_1}^{c_2}) = \frac{\epsilon}{(p+1)^o} \prod_{j=1}^{o} \sum_{i=0}^{p} t(\langle \bar{f}_{d_1}^{d_2}, j \rangle \mid \langle \bar{e}_{c_1}^{c_2}, i \rangle),$$

where p and o are the lengths in words of the sequences $\bar{e}_{c_1}^{c_2}$ and $\bar{f}_{d_1}^{d_2}$, respectively. $\langle \bar{f}_{d_1}^{d_2}, j \rangle$ is the j-th word of the sequence $\bar{f}_{d_1}^{d_2}$, and $\langle \bar{e}_{c_1}^{c_2}, i \rangle$ is the i-th word of the sequence $\bar{e}_{c_1}^{c_2}$. $t(\langle \bar{f}_{d_1}^{d_2}, j \rangle \mid \langle \bar{e}_{c_1}^{c_2}, i \rangle)$ is a statistical translation dictionary which stores the probability of the target word $\langle \bar{e}_{c_1}^{c_2}, i \rangle$ being translated into the source word $\langle \bar{f}_{d_1}^{d_2}, j \rangle$. This dictionary can be estimated automatically from the bilingual corpus by using the estimation methods described in (Brown et al., 1993). The software used to obtain this statistical dictionary was GIZA++ (Och and Ney, 2000; Knight, 1999). The probability that the sequences of words of the pair have a certain length (number of words) is measured by the ϵ term .

Now, expanding expression (4) with (5), we have:

$$\Pr(\bar{f}_1^b \mid \bar{e}_1^a) = \sum_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} Pr(\mathbf{s} \mid \bar{e}_1^a) \prod_{q=1}^{|\mathbf{S}|} \Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}). \quad (6)$$

However, we are interested in computing only the best segmentation, so, we define the most probable segmentation probability, $\widehat{\Pr(\bar{f}_1^b \mid \bar{e}_1^a)}$, as the maximum of expression (6):

$$\widehat{\Pr}(\bar{f}_1^b \mid \bar{e}_1^a) = \max_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} \Pr(\mathbf{s} \mid \bar{e}_1^a) \prod_{q=1}^{|\mathbf{S}|} \Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}).$$

Considering $\Pr(\mathbf{s} \mid \bar{e}_1^a)$ to be equally probable for all $\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)$ (that is, $C = \Pr(\mathbf{s} \mid \bar{e}_1^a)$) and assuming that $\Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q})$ is computed using the Model 1:

$$\widehat{\Pr}(\bar{f}_1^b \mid \bar{e}_1^a) = C \cdot \max_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} \prod_{q=1}^{|\mathbf{S}|} M_1(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}). \quad (7)$$

In order to obtain the segmentation with maximum probability, we want the argument that maximizes expression (7), so, we look for:

$$\hat{\mathbf{s}} = \operatorname*{argmax}_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} \prod_{q=1}^{|\mathbf{S}|} M_1(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}). \quad (8)$$

To solve the maximization problems (7) and (8), we use a dynamic programming scheme. In order to reduce the computational search cost, we impose a new restriction: no more than k initial segments can be joined for $\bar{f}_1^b$ or $\bar{e}_1^a$.

The algorithm to compute the probability of the best segmentation uses a bidimensional matrix

$\mathbf{s}[d,c]$. A graphical representation of this structure is shown in Figure 1, where the rows correspond to the initial segments of the source sentence and the columns correspond to the initial segments of the target sentence.

The expression which is computed for every position of the matrix $\mathbf{s}$ in Figure 1 is:

$$\mathbf{s}[d,c] = \max_{\substack{i=0..k \\ j=0..k}} \mathbf{s}[d-j-1, c-i-1] \cdot M_1(\bar{f}_{d-j}^{d} \mid \bar{e}_{c-i}^{c})$$

$\mathbf{s}[d,c]$ is the probability of translating the sequence of words $\bar{e}_1^c$ into the sequence of words $\bar{f}_1^d$, $\Pr(\bar{e}_1^c \mid \bar{f}_1^d)$.

The algorithm for computing the probability of the best parallel segmentation for an initial segmentation based on anchor words is shown in algorithm 1. When the computation of every $\mathbf{s}[d,c]$ is done, the probability of the best segmentation is stored in $\mathbf{s}[b,a]$

$$\begin{array}{c|c||c|c|c|c|c|}
 & \multicolumn{6}{c}{c = 1 \ldots a} \\
 & & & \bar{e}_1 & \bar{e}_2 & \ldots & \bar{e}_a \\
\hline
 & & 1 & 0 & 0 & \ldots & 0 \\
 & \bar{f}_1 & 0 & & & & \\
d = 1 \ldots b & \bar{f}_2 & 0 & & & & \\
 & \vdots & \vdots & & & & \vdots \\
 & \bar{f}_b & 0 & & & \ldots & \\
\hline
\end{array}$$

Figure 1: Graphical representation of the matrix $\mathbf{s}[d,c]$ used for the computation of $\widehat{\Pr}(\bar{f}_1^b \mid \bar{e}_1^a)$ in the dynamic programming algorithm 1.

Another matrix can be computed together with the matrix $\mathbf{s}$ in order to store the path for the most probable segmentation, that is, to store the groupings of initial segments that are carried out for the most probable segmentation.

2.3 A complete example

Now we offer a complete example of the computation of the segmentation of a pair of sentences. This pair of sentences is extracted from the FUB corpus (Vidal, 2000). This corpus is a bilingual text corpus of Italian-English pairs of sentences with restricted semantic domain. The sentences in the corpus are typical sentences of a tourist in the hotel domain, for example:

- *A che ora é disponibile il servizio navetta per l'aeroporto? / At what time is the shuttle service to the airport available?.*
- *Avete una stanza libera dal quattro al dieci Settembre? / Do you have a free room from the fourth to the tenth of September?.*

Defining the following sets of anchor words for Italian and English, respectively:

Italian Anchors
., ,, :, ;, ?, !, con, per, e, perché, vorrei, volevo

English Anchors
., ,, :, ;, ?, !, with, for, and, because, *I would like, I wish*

The English expressions *I would like* and *I wish* were treated as atomic anchor words.

This is a pair of sentences extracted from the corpus:

> *buonasera , sono la signora Rossi della camera trecentodue , vorrei disdire per domani mattina la colazione in camera , grazie .*
>
> *good evening , it is Mrs Rossi from room three hundred and two , I would like to cancel breakfast in room for tomorrow morning , thanks .*

The initial segmentation for the original sentences and the anchor words is shown in Figure 2.

After running Algorithm 1 described in section 2 on the initial segmentation of Figure 2, we obtained the segmentation shown in Figure 3 as the best segmentation.

3 Experiments

3.1 Corpora description

The EuTrans-I corpus (Vidal, 2000) is a Spanish-English corpus which was generated semi-automatically for the EuTrans-I task which is a subtask of the "Traveler Task". The

Algorithm 1: Algorithm for the computation of the probability of the best parallel segmentation for an initial segmentation based on anchor words $(\bar{e}_1^a, \bar{f}_1^b)$.

```
INPUT: $(\bar{e}_1^a, \bar{f}_1^b)$: initial segmentation;
        $k$: maximum number of consecutive initial segments that can be joined;

OUTPUT: $\widehat{\Pr}(\bar{f}_1^b \mid \bar{e}_1^a) \equiv$ probability of the best parallel segmentation for $(\bar{e}_1^a, \bar{f}_1^b)$;

VAR: $\mathbf{s}$: matrix to compute the best probability;

BEGIN
for ($c$=1; $c$ <=$a$; $c$++)                  /* For every initial segment in $\bar{e}$ */
    for ($d$=1; $d$ <=$b$; $d$++)               /* For every initial segment in $\bar{f}$ */
        {
        $\mathbf{s}[d, c] = 0.0$;
        /* Try to join $\bar{e}_c$ with previous initial segments: $\bar{e}_{c-1} \ldots \bar{e}_{c-k}$ */
        for($i$=0; $i$ <=$k$; $i$++)
            /* Try to join $\bar{f}_d$ with previous initial segments:$\bar{f}_{d-1} \ldots \bar{f}_{d-k}$ */
            for($j$=0; $j$ <=$k$; $j$++)
                {
                /* Store the best probability */
                $aux = \mathbf{s}[d-j-1, c-i-1] \cdot M_1(\bar{f}_{d-j}^d \mid \bar{e}_{c-i}^c)$;
                if ($aux > \mathbf{s}[d, c]$)    $\mathbf{s}[d, c] = aux$;
                }
        }
return($\mathbf{s}[b, a]$);
END
```

domain of the corpus is a human-to-human communication situation at a reception desk of a hotel. The corpus characteristics are shown in Table 1.

The FUB corpus (Vidal, 2000), is a bilingual Italian-English corpus with a restricted semantic domain. The application is the translation of queries, requests and complaints that a tourist can make at the front desk of a hotel, for example, asking for a booked room, requesting a service of the hotel, etc. The characteristics of the corpus are shown in Table 2.

3.2 Results

There is no standard method for evaluating the quality of a segmentation. One possible method is to compare the segmentation produced by the approach presented here with respect to a reference segmentation produced by hand. However, this is a very expensive procedure which is not error free. Another possible method for assessing the performance of this new segmentation technique is to compare the efficiency of a translation system obtained from the original corpus and another obtained from the segmented corpus on the translations of a test set of sentences.

We trained two finite-state transducers: one from the original parallel corpus and one from the segmented parallel corpus. In order to infer the transducers from a parallel corpus we used a technique known as Grammatical Inference and Alignments for Transducer Inference (GIATI) (Casacuberta, 2000). The translation quality was measured for every transducer on the test set by using the translation word error rate (TWER). This is the average number of wrong words in the translations generated by the transducer with respect to fixed reference translations for the source sentences.

ITALIAN INITIAL SEGMENTS
buonasera ,
sono la signora Rossi della camera trecentodue ,
vorrei
disdire per
domani mattina la colazione in camera ,
grazie .

ENGLISH INITIAL SEGMENTS
good evening ,
it is Mrs Rossi from room three hundred and
two ,
I would like
to cancel breakfast in room for
tomorrow morning ,
thanks .

Figure 2: Initial segmentation from the original sentences of the FUB corpus and the sets of anchor words.

ITALIAN	ENGLISH
buonasera ,	*good evening ,*
sono la signora Rossi della camera trecentodue ,	*it is Mrs Rossi from room three hundred and two ,*
vorrei	*I would like*
disdire per domani mattina la colazione in camera ,	*to cancel breakfast in room for tomorrow morning ,*
grazie .	*thanks .*

Figure 3: Final parallel segmentation for the example pair of sentences of the FUB corpus.

Table 1: Training and test data sets of the bilingual corpus EuTrans-I.

Training

	Spanish	English
N. Sentences	10,000	
N. Words	97,131	99,292
Vocabulary size	686	513
Perplexity(bigram)	8.6	5.2

Test

	Spanish	English
N. Sentences	2,996	
N. Words	35,023	35,590
Vocabulary size	613	469

Table 2: Training and test data sets of the bilingual FUB corpus 5.1.

Training

	Italian	English
N. Sentences	3,038	
N. Words	55,302	64,176
Vocabulary size	2,459	1,712
Perplexity(bigram)	31	25

Test

	Italian	English
N. Sentences	300	
N. Words	6,121	7,243
Vocabulary size	715	547

The number of initial segments that were allowed to be joined in one segment of the final segmentation was five.

In order to infer a finite-state transducer, the GIATI technique needs word-level alignments such as those described in (Brown et al., 1993; Knight, 1999) for every pair of sentences of the training set. Model 4 (Brown et al., 1993) was estimated with the non-segmented corpus and word alignments were obtained. With the segmented corpus, each pair of segments was considered as a pair of sentences, Model 4 was estimated and the corresponding word alignments were computed. These alignments were computed using the software GIZA++ (Och and Ney, 2000; Knight, 1999), obtaining the alignments produced by Model 4 (Brown et al., 1993). The finite-state transducer generated with GIATI is derived from a n-gram model inferred from the source sentences. In these source sentences, the words of every input sentence are labeled with the words of the corresponding target sentence according to the word alignments obtained with Model 4.

Tables 3 and 4 show the average lengths of the source-target sentences, along with the lengths of the segmented sentences obtained by the proposed technique. It is worth noting that on the average, the more complex and long sentences of the FUB corpus are broken down into much shorter (and simpler) segments.

Table 3: Average sentence length (number of words) for the EUTRANS-I training set in the non-segmented and segmented versions.

	Spanish	English
Non-segmented	9.71	9.93
Segmented	7.40	7.57

Table 4: Average sentence length (number of words) for the FUB training set in the non-segmented and segmented versions.

	Italian	English
Non-segmented	17.94	21.55
Segmented	4.79	5.76

Table 5 shows the TWER values for the inferred transducers from the EUTRANS-I training set using the Model 4 alignments and fourgrams for GIATI. Table 6 shows the TWER values for the corresponding transducers of the FUB training set using the Model 4 alignments and bigrams for GIATI.

The transducer inferred using the segmented EUTRANS-I corpus produced a greater error rate than the transducer inferred using the non-segmented corpus. On the other hand, the results for the segmented FUB corpus improved the results over those obtained for the non-segmented version of the corpus.

Table 5: TWER for the EUTRANS-I test set using the transducers inferred with GIATI using fourgrams and the Model 4 alignments.

Non-segmented	8.0
Segmented	10.5

Table 6: TWER for the FUB test set using the transducers inferred with GIATI using bigrams and the Model 4 alignments.

Non-segmented	26.6
Segmented	25.2

4 Conclusions

A new automatic segmentation technique for a parallel corpus has been presented. The method has been tested using the translation results obtained for two tasks: the EUTRANS-I task and the FUB task.

The EUTRANS-I task is relatively much simpler than the FUB task, and the length of the sentences is significantly shorter. Consequently, alignment models such as Model 4 produce very good results on *unsegmented* pairs of this corpus, thereby directly leading to good translation results with GIATI transducers trained on unsegmented aligned data. The FUB corpus, on the other hand, is much more complex and the lengths of the sentences are much longer. For these (long) pairs of sentences, alignments obtained by alignments models such as Model 4 tend not to be as good as those of EUTRANS-I. In this case, using the shorter pairs of sentences obtained by the proposed segmentation technique definitely helps the alignment model to produce better alignments, thereby leading to improved results for the GIATI transducers trained on *segmented* aligned pairs. It should be noted that the FUB task is much more realistic than the EUTRANS-I task.

Although the proposed technique has a heuristic component (the selection of the anchor words sets), it improves the translation results with minimum human effort, especially for difficult tasks such as the FUB task.

References

P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J. Lafferty, R.L. Mercer, and P. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–310.

F. Casacuberta. 2000. Inference of finite-state transducers by using regular grammars and morphisms. In *Proceedings of International Conference on Grammatical Inference - ICGI2000*, pages 1–14.

K. Knight. 1999. A statistical MT tutorial workbook. Technical Report prepared in connection with the JHU summer workshop, Johns Hopkins Univ.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *ACL00*, pages 440–447, Hongkong, China, October.

E. Vidal. 1997. Finite-state speech-to-speech translation. In *Proceedings of the International Conference on Acoustic, Speech and Signal Processing*, volume 1, pages 111–114.

E. Vidal. 2000. Final report. Technical Report EuTrans project, Technical Report Deliverable D0.1c, Information Technology. Long Term Research Domain. Open Scheme. Project Number 32026.

Computer-based Support for Patients with Limited English

Harold Somers
Centre for Computational Linguistics
UMIST
Manchester
Harold.Somers@umist.ac.uk

Hermione Lovel
School of Primary Care
Faculty of Medicine
University of Manchester
Hermione.Lovel@man.ac

Abstract

The paper describes a proposal for computer-based aids for patients with limited or no English. The paper describes the barriers to health-care experienced due to linguistic problems, then suggests some computer-based remedies incorporating a multi-engine machine translation system based on a corpus of doctor–patient interviews which provides a dialogue model for the system. The doctor's and patient's interfaces are described. Ideas from Augmentative and Alternative Communication and in particular picture-based communication are incorporated. The initial proposal will focus on Urdu- and Somali-speaking patients with respiratory problems.

1 Introduction

This paper describes a proposed framework for the development of computer-based aids for patients with limited or no English. Aimed at users of the Health Services who are disadvantaged by their (lack of) linguistic skills, the system will assist the patient in different ways at different stages of their interactions with health-care providers. In its full conception it will embrace a wide range of NLP technologies.

Focusing on the GP's clinic, it will provide a kind of FAQ help-desk and act as a kind of Receptionist to help determine whether the patient needs to see the GP or some other health-care specialist. If a GP consultation is indicated, the computer can be used for history note-taking. During the consultation itself, it can act as a mediator between the doctor and patient. Afterwards, in help-desk mode again, it can help the patient understand the diagnosis, any tests needed, and the proposed treatment regime.

We propose in the first instance to develop systems aimed at Urdu- and Somali-speaking patients, focusing on respiratory problems (e.g. asthma).

2 Patients with Limited English

In many parts of the UK there are recent or long-term immigrants, refugees, and asylum seekers and other people whose command of English, while often adequate for day-to-day activities such as shopping and other domestic chores, is not sufficient for more formal situations such as interactions with health services, especially visits to their GP. There is no shortage of literature reporting disparities in health, health-care, and social care provision in these communities and communication difficulties are identified as a major factor [1]. The problem is also well recognised in other countries [2].

People in this situation will only rarely be lucky enough to find a homolingual GP, which is probably the preferred option [3] or (less than ideal) an interpreter or linkworker, who may also

[1] e.g. McAvoy and Sayeed (1990), Chalabian and Dunnington (1997), Acheson (1998), Smith (2000), Woodhead (2000), Burnett and Peel (2001)

[2] e.g. USA (Uba 1992; Hornberger et al 1996; Jackson 1998), Canada (Fowler 1998), New Zealand (Blakely 1996), Australia (Sinnerbrink et al. 1996; Silove et al. 1999; Nerad et al. 2000), Norway (Karlsen et al. 1998), Sweden (Sundquist et al. 1999), Austria (Pöchhacker 2000), Switzerland (Blöchliger et al. 1997; Graz et al. 2002), Ivory Coast (Zotti 1999)

[3] Bhui (1998)

have varying appropriateness [4]. Even then they will still have to communicate with other persons (receptionist [5], community nurse, pharmacist, specialist). Some may take with them an "interpreter", typically a family member (including an inappropriate child [6]) or someone from their religious community, or else will just "muddle through" (with both clients and providers often using ingenious ways to express themselves) [7]. The outcome is undesirable in either case, for numerous reasons. In recent systematic literature searches of a range of medical and social science journal databases since 1990, on barriers to accessing health-care experienced by refugees in the UK [8], language difficulties were identified as the largest single barrier to care and as such repeatedly identified as a major concern for refugees [9]. In a study in London [10], 53% of GPs felt that language difficulties were a problem. A survey of the Vietnamese community in Greenwich [11] revealed that 17% of respondents had changed their minds about visiting the GP because of lack of access to an interpreter. Effective communication is important in all areas of health care [12], from finding out about services available through to complying with treatment.

There have been only a few suggestions for initiatives to tackle this problem [13], including a cheap national specialist medical telephone interpreting service, with hands-free conferencing to enable concurrent discussions and examination if needed [14], use of the Red Cross multilingual phrasebook [15], and multilingual phrase cards for use by health-care practitioners and receptionists (simple words like days of the week could make a significant difference to people trying to access health care). Further initiatives urgently need to be developed.

There can be no doubting the importance of doctor–patient communication, which has for many years been the focus of medical attention. Everything in medical practice arguably derives from the consultation, during which the doctor must acquire and impart information, and set up a relationship with the patient; the consultation itself can also have a therapeutic role. Valuable consultation time may be saved by having the patient complete a pre-consultation questionnaire which allows information to be expressed which may be given reluctantly in a hurried interview. There is a considerable literature on the structure of the consultation, from various angles including the linguistic, pragmatic, ergonomic, social and of course medical aspects. Effective communication improves outcomes [16] and it is argued [17] that doctors have responsibilities to their patients that can only be met by effective communication

Use of computers in the doctor–patient consultation paradoxically has been recognised as both potentially detrimental and potentially hugely helpful. The early use of computers on the consultation desk was seen as a threat, detracting from interaction with the patient, reducing eye contact and rapport build up. More recently the help of computers to increase communication and rapport has begun to be recognised. Computers can help in accessing records of other 6-minute interactions, reducing the need for repetition. A recent systematic review of UK literature in the 1990s [18], described as rich in description but low on evaluative information, did conclude that Primary Care computing systems can improve practitioner performance, particularly for health promotion interventions. It also reported that this may be at the expense of patient-initiated activities, and that many practitioners are suspicious of the negative impact on relationships with patients. The review showed that there remains a dearth of evidence evaluating effects of use of computers on patient outcomes.

3 A Computer-based Solution

As mentioned above, the proposed system will operate in various "modes". The most intricate of these is during the consultation itself, when it will serve as a kind of interactive phrase-book, designed to run on the typical PC that might be found on a GP's desk.

[4] Phelan and Parkman (1995), Gillam and Levenson (1999).
[5] Free (1998).
[6] Jones and Gill (1998), Burnett and Peel (2001b)
[7] Montgomery (2000)
[8] Jary (2001), Hays (2002)
[9] Lam and Green (1994), Tang and Cuninghame (1994),
[10] Ramsay and Turner (1993)
[11] Lam and Green (1994)
[12] Voelker (1995)
[13] Reviewed in Jary (2001)
[14] Wolmuth (1996), Jones and Gill (1998a,b)
[15] Matthews (1999)
[16] Stewart (1995)
[17] Meryn (1998)
[18] Mitchell and Sullivan (2001)

At the core of the system is a hybrid multi-engine embedded MT system: essentially an EBMT system with a "translation memory" (TM) extracted from corpora of doctor–patient interviews, supplemented with a simple rule-based MT (RBMT) system and a word-by-word lexical look-up facility. It will have a highly flexible interface: a simple set-up like in a chat-room, where each user types at a keyboard with the results shown on a split-screen is not practical when one of the users may not be a regular computer user.

The system in this mode has two users: the doctor and the patient, with significantly different profiles of computing experience. Accordingly, the user-interfaces will be quite different for the two users, while necessarily being integrated. Whereas the doctor can be expected to use the keyboard and mouse, and be comfortable with a sophisticated GUI, the patient's interface presents a number of problems.

Obviously, in the long-term we would want to consider speech input and output for both the doctor's and patient's interfaces. In the short term, and given the current state-of-the-art, text-based interfaces are proposed.

It should also be remembered that some patients will not need to use the system for every part of the interview, their English being sufficient for some interactions. In addition to the "Consultation mode", we will simultaneously develop a "Reception mode" with an interactive FAQ/help system and a "History mode" involving a computer-aided patient interview system.

In the following sections, we give some more details about the design features of the different modes of the proposed system.

3.1 Multi-engine MT system

MT has now proved itself viable under conditions of restricted input and interactive use. Particularly effective is an architecture which tries various strategies in parallel and then tries to reconcile the results. This is the "multi-engine" approach seen in the PANGLOSS and DIPLOMAT systems [19]. The engines that our system will use will be an EBMT/TM system, a rule-based transfer system, and a simple lexical look-up system; it is to be expected that the input from the doctor will usually go through the EBMT system, while the patient's input, being more varied, may more often be translated by RBMT or on a word-by-word basis. In the proposed scenario, it is an example of an "embedded" MT system [20].

EBMT is akin to case-based reasoning (CBR) [21] in that new translations are composed on the basis of past translations, as provided by the "example base" of utterances taken from a corpus of doctor–patient interviews, manually translated into the target language. This method gives a very high quality of translation when the input can be matched against an appropriate example. The match does not have to be exact: as in CBR, a partial match can lead to a successful outcome.

RBMT and word-by-word translation methods tend to result in more stilted translations, closely following the syntax of the source language. In our scenario, this is more likely to be used for translating the patient's replies into English: thus the burden of understanding a less polished translation will normally fall on the doctor, who will gain experience of the system with use, and – on the evidence of early users of less sophisticated MT systems [22] – will quickly get used to its quirky style.

The notion of "restricted input" relates to the widely accepted notion of "sublanguage"-based approaches to MT [23], especially inasmuch as a corpus can help to define the sublanguage [24].

The experience of the DIPLOMAT project is especially relevant to this proposal, since their system was developed specifically with rapid development of new language pairs for use in a dialogue situation between an experienced user and a naïve interviewee who may have little experience of computers, and may not even be literate. Versions of DIPLOMAT have been developed for English–Croatian and English–Haitian Creole, for use in the field to allow English-speaking soldiers on peace-keeping missions to interview local residents [25]. An additional feature of DIPLOMAT is the use of speech-recognition and synthesis front and back ends, and the extensive use of on-screen

[19] Frederking et al. (1994, 1997)

[20] Van Ess-Dykema et al. (2000)

[21] See Somers and Collins (2003)

[22] cf. Church and Hovy (1993)

[23] Kittredge and Lehrberger (1982)

[24] cf. Deville and Herbigniaux (1995), McEnery and Wilson (1996:147ff), Sekine (1997)

[25] See also www.avt-actii.lmowego.com/

interactive correction by both participants. As the language pairs indicate, it has been tested in the former Yugoslavia, and in Haiti. The success of the DIPLOMAT project gives a strong indication of the viability of the current project.

3.2 Corpus of doctor–patient interviews

Transcribed corpus data from doctor–patient interviews is readily available in the British National Corpus, which contains about 100 examples of short (300–900 words) medical consultations in GP surgeries or hospitals, already annotated for POS tags and some other aspects. Several other similar corpora have been collected [26]. Other researchers have collections of tape-recordings [27], and there are even conferences dedicated to the analysis of doctor–patient discourse [28]. Data from consultations where an interpreter was present may also be relevant [29].

This corpus will serve multiple purposes, and accordingly we should distinguish various of its characteristics. For example, transcriptions of interpreter-mediated interviews, and interviews where the patient has a poor command of English, will be useful as an indication of how such interviews tend to proceed. They will not however serve as a direct model for the system, which aims to bypass some of the difficulties that arise in such situations. For most of our purposes, what is important is not so much the verbatim transcripts, but the model of the discourse and the examples of the kinds of things that are said [30]. This being the case, the utterances in the corpus can legitimately be “cleaned up”. The corpus will be marked up, especially for dialogue function in a TEI-conformant manner.

Another purpose of the corpus is to provide a source of examples for the EBMT system, and so a parallel target version will have to be provided. It will also serve as a training corpus for the development of the translation lexicon and the RBMT system. To some extent, some of this linguistic information can be extracted semi-automatically [31]. Finally, it can serve as a dialogue model, simplifying and determining the options offered in the menu-driven mode for both doctor and patient [32].

3.3. The doctor’s interface

Doctors greet and observe patients in all doctor–patient encounters, and in the UK the consultation proceeds normally these days in the presence of a computer which is used for recording all personal details, history taking of a problem, diagnosis, and treatment. Thus it is a small step to consider the possibility of using a computer to aid communication as part of the existing situation.

For the doctor’s interface, two main possibilities are envisaged: typing at the keyboard, augmented by auto-completion; and a menu-based approach, enriched by dynamic domain knowledge.

The menu-based interface, which is also appropriate for the patient’s interface, involves “intelligent” menu-driven selection. Several script- or frame-based interfaces have been reported, for example the UNICORN system [33], which is specifically aimed at multilingual communication, DRAFTER [34] for multilingual document preparation, Floorgrabber [35] and Frametalker [36] for users with communication difficulties. The “intelligence” derives from domain knowledge and a discourse model which permit the interface to be simplified by determining the options offered. This type of interface is most appropriate when the consultation is following a predictable course, and “standard” questions or comments are being made, for example “How long have you had this problem?”.

In the keyboard-based typing interface, the doctor simply types the input, or parts of it that the patient does not understand. Typing is aided by auto-completion proposals based on the corpus, an idea already demonstrated in the TRANSTYPE project [37]. Typing is necessitated when what the doctor wants to say is not sufficiently similar to anything that the menu-driven interface is offering,

[26] For example by Thomas and Wilson (1996) and Wynn (1999)

[27] See for example ww2.mcgill.ca/ Psychiatry/ transcultural/ prmary.html

[28] For example, the Conference on Medical Interaction, 18-20 October 2000, at the University of Southern Denmark, Odense. See www.conversationanalysis.net/Conferences/Medcal/program_doc-pat.htm.

[29] Cambridge (1997)

[30] cf. Passonneau and Litman (1997), Berthelin et al. (1999)

[31] See for example Brent (1993), Smadja (1993), Melamed (2000), Véronis (2000)

[32] cf. Alm et al. (1989)

[33] Dye et al. (1997), Iwabuchi et al. (2000)

[34] Hartley and Paris (1997)

[35] Alm and Arnott (1998)

[36] Higginbotham et al. (2000)

[37] Langlais et al. (2000)

for example a much more specific question or comment which relates to things the patient may have said earlier, e.g. “When did your step-mother pass away?”.

3.4 The patient’s interface

Some patients will be highly experienced in using computers while for others, a keyboard- or mouse-driven interface may not be appropriate. Therefore, a range of interfaces must be made available to the patient. We can include simple interfaces like a drop-down menu, as in the doctor’s interface. If the patient’s language involves a different character set (as is the case with Urdu), it is not viable to assume the patient might want to use the keyboard: character-handling of non-Roman writing systems is not a problem as such (and is necessary for output), but we cannot assume that the patient can quickly learn to use an Urdu keyboard, or, worse still, to learn a set of mappings from a QWERTY keyboard. The problem may be less acute for Somali-speaking patients, whose language is written using the Latin alphabet on a straightforwardly phonemic basis. All these issues represent an important and innovative aspect of the research proposed here: we need to discover the best way to integrate all the possibilities so as to provide an interface that both doctor and patient are comfortable with, that promotes an equitable exchange (rather than giving one or other user excessive control), and makes best use of their respective skills and experience. There are important socio-cultural issues here which we cannot address fully in this paper

Of relevance here is the field of Augmentative and Alternative Communication (AAC) and in particular the work on picture-based communication (PBC) interfaces [38]. AAC is usually focused on disabled users, and AAC techniques have apparently not been applied to users whose only “handicap” is lack of a shared language [39]. Langer and Hickey (1999) report on growing There are growing contacts between the AAC and NLP research communities [40]. One group [41] developed a GUI for healthcare workers in rural India, like us facing the problems of inexperienced computer users and a non-Roman writing system. HCI issues are of paramount importance here: robustness and flexibility are essential; alternative modes of input, such as touch screens, may be preferred, since the patient may lack experience of mouse manipulation.

3.5 “Reception mode”: FAQ/Help desk

Consultations often include obtaining answers to the same series of questions (such as how long has the problem been continuing). This may lend itself to identification of a series of frequently asked questions in the form of a pre-consultation computer-mediated help-desk and interview [42]. By “help desk”, we mean a simple on-line interface containing potted texts in answer to frequently asked questions (FAQs).

These interfaces can be run with a simulated natural-language interface based on key-word matching. This could be installed on a computer terminal in the Health Centre reception area, so that potential patients could get relevant information without even making an appointment with the GP. There has been a considerable amount of relevant work in this area, notably on Tailored Patient Information (TPI) systems [43]. Navigation of the help facility can be system-led or patient-led. In the latter case it would work in much the same way as the help facility in, say, a word-processor offers “Type in your query here”. In the former case, the user is lead through the interaction with a structured database depending on the choices made at each point. Different start points might relate to basic symptoms (answering the question “Do I need to see the doctor?”), general procedure (“What can I expect when I go to the hospital?”) or, after diagnosis, what the course of treatment involves, e.g. general information about the drugs or therapy that have been prescribed, and the likely outcomes and progress of the patient’s condition.

[38] Blenkhorn (1992), Loncke et al. (1999)

[39] Personal communication: Pat Mirenda, editor of the journal *AAC Augmentative and Alternative Communication*. See also Johnston (in prep.).

[40] Copestake et al. (1997), Langer (1998)., Langer and Hickey (1999),

[41] Grisedale et al. (1997)

[42] cf. Osman et al. (1994)

[43] Buchanan et al. (1995), Cawsey et al. (1995), Reiter and Osman (1997)

3.6 "History mode": Computer-mediated interviewing

Many services in general are finding it helpful nowadays to gather basic information from the patient prior to meeting with the professional. This is the important element of "history" note taking which can be partly accomplished using computer-mediated interviewing techniques, which can make better use of the time the patient spends in the waiting room. These widely-used techniques have been found to be particularly useful in sensitive applications like taking patient's medical details [44], where decreased time pressure leads to fuller responses, especially when questions are of a sensitive or embarrassing nature. Most systems are based on flexible multiple-choice questionnaires, while the use of free text [45] is more complex, and brings us into the area of conversation systems. An on-line consultation might be appropriate in the case of patients returning with chronic problems.

4 Conclusion

We have presented here a proposal for a highly innovative multi-modal system. While plan-based communication or authoring tools have been proposed previously, the multilingual profile coupled with the dialogue situation for the doctor's and patient's interfaces is quite novel. The application of AAC techniques to use by non-handicapped but linguistically disadvantaged users is likewise a new idea. This presentation has focused on the language technology aspects, but the work has a simultaneous impact for researchers in primary care, implying research on doctor–patient communication, access to health services by, and improving the quality of access and quality of care to hard-to-reach groups [46], reducing perceived time wasting with perceived difficult patients, developing training agendas for health care professionals, and agendas for community development initiatives [47] so that newly arrived communities make better use of the local health services and get a better quality of care not only in the UK but in other countries across Europe, Australasia and North America. It is at the moment a proposal, but we hope in due course to be able to report on its implementation, and on results of trials and evaluations.

References

Acheson, D. 1998. *Independent enquiry into inequalities in health.* London: Stationery Office.

Alm, N. and J. L. Arnott. 1998. 'Computer-assisted conversation for nonvocal people using prestored texts'. *IEEE Transactions on Systems, Man, and Cybernetics* **28**, 318–328.

Alm, N., J. L. Arnott and A. F. Newell. 1989. 'Discourse analysis and pragmatics in the design of a conversation prosthesis'. *Journal of Medical Engineering and Technology* **13**, 10–12.

Berthelin, J.-B., B. Grau, I. Robba and A. Vilnat. 1999. 'A cross-corpus vision of conversation and dialogue'. *Language Technologies – Multilingual Aspects: Workshop in the framework of the 32nd Annual Meeting of the Societas Linguistica Europaea*, Ljubljana, www.limsi.fr/Individu/ jbb/cross-corpus.html.

Bhui, K.1998. 'The public favours bilingual staff over interpreters'. *Britsh Medical Journal* **317**, 816.

Blakely, T. 1996. 'Health needs of Cambodian and Vietnamese refugees in *Porirun*'. *New Zealand Medical Journal* **109**, 381–384.

Blenkhorn, P. 1992. 'A picture communicator for symbol users and/or speech impaired people'. *Journal of Medical Engineering and Technology* **16**, 243–249.

Blöchliger, C., M. Tanner, C. Halz, and T. Junghanss. 1997. 'Asylsuchende und Flüchtlinge in der ambulanten Gesundheitsversorgung: Kommunikation zwischen Arzt und Patient'. *Praxis: Schweizerische Rundschau für Medizin* **86**, 800–810.

Brent, M. R. 1993. 'From grammar to lexicon: unsupervised learning of lexical syntax'. *Computational Linguistics* **19**, 243–262.

Buchanan, B., J. Moore, D. Forsythe, G. Garenini, G. Banks and S. Ohlsson. 1995. 'An intelligent interactive system for delivering individualized information to patients'. *Artificial Intelligence in Medicine* **7**, 117–154.

Burnett, A. and M. Peel. 2001a. 'Asylum seekers in Britain: What brings asylum seekers to the UK?' *British Medical Journal* **322**, 485–488.

Burnett, A. and M. Peel. 2001b. 'Asylum seekers in Britain: Health needs of asylum seekers and refugees'. *British Medical Journal* **322**, 544–547.

Cambridge, J. 1997. *Information Exchange in Bilingual Medical Interviews.* MA dissertation, Department of Linguistics, University of Manchester.

Cawsey, A., K. Binsted and R. Jones. 1995. 'Personalised explanations for patient education'. *Proceedings of the 5th European Workshop on Natural Language Generation*, Leiden, Netherlands, pp. 59–74.

[44] Lilford et al. (1985)

[45] For example Peiris et al. (1995)

[46] Lovel et al. (1998)

[47] Moran et al. (2000)

Chalabian, J. and G. Dunnington. 1997. 'Impact of language barrier on quality of patient care, resident stress, and teaching'. *Teaching and Learning in Medicine* **9**, 84–90.

Church, K. W. and E. H. Hovy. 1993. 'Good applications for crummy machine translation'. *Machine Translation* **8**, 239–258.

Copestake, A., S. Langer and S. Palazuelos-Cagigas (eds) 1997. *Natural Language Processing for Communication Aids: Proceedings of a workshop sponsored by the Association for Computational Linguistics*, Madrid.

Deville, G. and E. Herbigniaux. 1995. 'Natural language modelling in a Machine Translation prototype for healthcare applications: a sublanguage approach'. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, TMI* 95, Leuven, Belgium, pp. 142–157.

Dye, R., N. Alm, J. L. Arnott, G. Harper and A. I. Morrison. 1997. 'A script-based AAC system for transactional interaction'. *Natural Language Engineering* **1**, 1–13.

Fowler, N. 1998. 'Providing primary health care to immigrants and refugees: the North Hamilton experience'. *Canadian Medical Association Journal* **159**, 388–391.

Frederking, R., S. Nirenburg, D. Farwell, S. Helmreich, E. Hovy, K. Knight, S. Beale, C. Domashnev, D. Attardo, D. Grannes and R. Brown. 1994. 'Integrating translations from multiple sources within the Pangloss Mark III Machine Translation system'. *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, pp. 73–80.

Frederking, R., A. Rudnicky and C. Hogan. 1997. 'Interactive speech translation in the DIPLOMAT project'. *Spoken Language Translation: Proceedings of a Workshop Sponsored by the Association for Computational Linguistics and by the European Network in Language and Speech (ELSNET)*, Madrid, pp. 61–66.

Free, C. 1998. 'Some ethnic groups may have problems getting as far as a consultation'. *British Medical Journal* **317**, 816.

Gillam, S. and R. Levenson. 1999. 'Linkworkers in primary care'. *British Medical Journal* **319**, 1215.

Graz, B., J.-P. Vader and M.-F. Raynault. 2002. 'Réfugiés, migrants, barrière de la langue: opinion des praticiens sur les moyens d'aide à la traduction'. *Santé Publique* **14**, 75–81.

Grisedale, S., M. Graves and A. Grünsteidl. 1997. 'Designing a graphical user interface for healthcare workers in rural India'. *Human Factors in Computing Systems: CHI 97 Conference Proceedings*, Atlanta, Georgia, pp. 471–478.

Hartley, A. and C. Paris. 1997. 'Multilingual document production from support for translating to support for authoring'. *Machine Translation* **12**, 109–128.

Hays, L. 2002. *Barriers to health care for refugees and asylum seekers.* Special Study Module, Medical Undergraduate Course, Manchester, Supervisor Dr Hermione Lovel.

Higginbotham, J., B. J. Moulton, G. W. Lesher, D. P. Wilkins and J. Cornish. 2000. 'Frametalker: development of a frame-based communication system'. *Proceedings of the CSUN Conference on Technology and Disability*, Los Angeles. Proceedings online www.csun.edu/cod/conf2000/ proceedings/0156Higginbotham.html.

Hornberger, J. C., C. D. Gibson Jr., W. Wood, C. Dequeldre, I. Corso, B. Palla and D. A. Block. 1996. 'Eliminating language barriers for non-English-speaking patients'. *Medical Care* **34**, 845–856.

Iwabuchi, M., N. Alm, P. Andreasen and K. Nakamura. 2000. 'The development of UNICORN – a multilingual communicator for people with cross-language communication difficulties'. *Proceedings of the CSUN Conference on Technology and Disability*, Los Angeles. Proceedings online www.csun.edu/cod/conf2000/proceedings/ 0025Alm.html.

Jackson, C. 1998. 'Medical interpretation, an essential service for non-English speaking immigrants'. In S. Loue (ed.) *Handbook of Immigrant Health*, London: Plenum Press, pp. 61–79.

Jary, D. 2001. *What are the barriers to accessing health care experienced by refugees in the United Kingdom?* Special Study Module, Medical Undergraduate Course, Manchester, Supervisor Dr Hermione Lovel.

Johnson, M. in prep. '*Disabled* by language? How can we *enable* ethnic minority patients with limited/no English to communicate with health-care professionals?' (Paper in prep. for submission to *Language and Communication*).

Jones, D. and P. Gill. 1998a. 'Breaking down language barriers, the NHS needs to provide accessible interpretimg services for all'. *British Medical Journal* **316**, 1476.

Jones, D. and P. Gill. 1998b. 'Refugees and primary care: tackling the inequalities'. *British Medical Journal* **317**, 1444–1446.

Karlsen, W. B. and A. L. Haabeth. 1998. 'Telefontolk: Et godt alternativ til tradisjonell tolkebruk.' *Tidsskrift for Den norske lægeforening* **118**, 253–254.

Kittredge, R. and J. Lehrberger (eds) 1982. *Sublanguage: Studies of Language in Restricted Semantic Domains*, Berlin: Walter de Gruyter.

Kolodner, J. 1993. *Case-Based Reasoning.* San Mateo, CA: Morgan Kaufmann.

Lam, T. and J. Green. 1994. 'Primary Care and the Vietnamese community: a survey in Greewich'. *Health and Social Care in the Community* **2**, 293–299.

Langer, S. (ed.) 1998. 'Special issue on Augmentative and Alternative Communication'. *Natural Language Processing* **4**.1.

Langer, S. and M. Hickey. 1999. 'Augmentative and Alternative Communication and Natural Language Processing: current research activities and prospects'. *AAC Augmentative and Alternative Communication* **15**, 260–268.

Langlais, P., G. Foster and G. Lapalme. 2000. 'Unit completion for a computer-aided translation typing system'. *Machine Translation* **15**, 267–294.

Lilford, R. J., P. Bingham, G. L. Bourne and T. Chard. 1985. 'Computerized histories facilitate patient care in a termination of pregnancy clinic: the use of a small computer to obtain and reproduce patient information'. *British Journal of Obstetrics and Gynaecology* **92**, 333–340.

Loncke, F. T., J. Clibbens, H. H. Arvidson and L. L. Lloyd. 1999. *Augmentative and Alternative Communication: New Directions in Research and Practice.* London: Whurr.

Lovel, H. J., Z. Mohamed and R. Moran. 1998. 'Needs assessment with a hard-to-reach group (the example of Somali refugees in Manchester)'. *Needs and Outcomes Assessment in Primary Health Care*, World Health Organisation, Geneva.

Lynch, M. A. and C. Cuninghame. 2000. 'Understanding the needs of young asylum seekers'. *Archives of Disease in Childhood* **83,** 384–387.

Matthews, P. 1999. 'Meeting health needs of asylum seekers; practical approaches can make care easier'. *British Medical Journal* **318**, 671.

McAvoy, B. and A. Sayeed. 1990. 'Communication'. In B.R. McAvoy and L.J. Donaldson (eds), *Health care for Asians*, Oxford: Oxford University Press, pp. 57–71.

McEnery, T. and A. Wilson. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Melamed, I. D. 2000. *Empirical Methods for Exploiting Parallel Texts.* Cambridge, Massachusetts: MIT Press.

Meryn, S. 1998. 'Improving doctor patient communication. Not an option, but a necessity'. *British Medical Journal* **316**, 1992.

Mitchell, E. and F. Sullivan. 2001. 'A descriptive feast but an evaluative famine: systematic review of published articles on primary care computing during 1980-97'. *British Medical Journal* **322**, 279–282.

Montgomery, S. 2000. 'Health care for asylum seekers: Main obstacles are inflexibility of NHS and bureaucracy of support systems'. *British Medical Journal* **321**, 893.

Moran, R., Z. Mohamed and H. J. Lovel. 2000. 'Taking action on refugee health needs'. *2nd International Conference on Women's Health*, Edinburgh.

Nerad, S. and A. Janczur. 2000. 'Primary Care with immigrant and refugee populations: Issues and challenges'. *Australian Journal of Primary Care-Interchange* **6**, 222–229.

Osman, L., M. Abdalla, J. Beattie, S. Ross, I. Russell, J. friend, J. Legge and J. Douglas. 1994. 'Reducing hospital admissions through computer supported education for asthma patients'. *British Medical Journal* **308**, 568–571.

Passonneau, R. J. and D. J. Litman. 1997. 'Discourse segmentation by human and automated means'. *Computational Linguistics* **23**, 103–139.

Peiris, D. R., N. Alm and P. Gregor. 1995. 'Computer interviews: an initial investigation using free text responses'. In M. A. R. Kirby, A. J. Dix and J. E. Finlay (eds) *People and Computers X: Proceedings of HCI '95, Huddersfield, ...*, Cambridge: Cambridge University Press, pp. 281–288.

Phelan, M. and S. Parkman. 1995. 'How to do it: work with an interpreter'. *British Medical Journal* **311**, 555–557

Pöchhaker, F. 2000. 'Language barriers in Vienna hospitals'. *Ethnicity & Health* **5**, 113–119.

Ramsay, R. and S. Turner. 1993. 'Refugees' health needs'. *British Journal of General Practice* **43**, 480–481.

Reiter, E. and L. Osman. 1997. 'Tailored patient information: some issues and questions'. *ACL Workshop From Research to Commercial Applications: Making NLP Technology Work in Practice*, Madrid, pp. 29–34.

Sekine, S. 1997. 'A new direction for sublanguage NLP'. In D. B. Jones and H. L. Somers (eds) *New Methods in Language Processing*, London: UCL Press, pp. 165–177.

Silove, D., Z. Steel, P. McGorry and J. Drobry. 1999. 'Problems Tamil asylum seekers encounter in accessing health and welfare services in Australia'. *Social Science and Medicine* **49**, 951–956.

Sinnerbrink, I., D. M. Silove, V. L. Manicavasgar, Z. Steel and A. Field. 1996. 'Asylum seekers: general health status and problems with access to health care'. *Medical Journal of Australia* **165**, 634–637.

Smadja, F. 1993. 'Retrieving collocations from text: Xtract'. *Computational Linguistics* **19**, 143–177

Smith, T. 2000. 'A refuge for children? The impact of the Immigration and Asylum Act'. *Poverty* **105**, 6–10.

Somers, H. and B. Collins. 2003. 'EBMT seen as case-based reasoning'. To appear in M. Carl and A. Way (eds) *Recent Advances in Example-Based Machine Translation*, Dordrecht: Kluwer Academic.

Stewart, M. A. 1995. 'Effective physician and patient communication and health outcomes: a review of the literature'. *Canadian Medical Association Journal* **152,** 1423–1433.

Sundquist, J., L. Bayard-Burfield, L. M. Johansson and S.-E. Johansson. 2000. 'Impact of ethnicity, violence and acculturation on displaced migrants: psychological distress and psychosomatic complaints among refugees in Sweden'. *Journal of Nervous and Mental Disease* **188**, 357–365.

Tang, M. and C. Cuninghame. 1994. 'Ways of saying'. *Health Service Journal* **104**, 28–30.

Thomas, J. and A. Wilson. 1996. 'Methodologies for studying a corpus of doctor-patient interaction'. In J. Thomas and M. Short (eds) *Using Corpora for Language Research*, London: Longman, pp. 92–109.

Uba L. 1992. 'Cultural barriers to healthcare for southeast Asian refugees'. *Public Health Reports* **107**, 545–548.

Van Ess-Dykema, C., C. R. Voss and F. Reeder (eds) *ANLP/NAACL 2000 Workshop: Embedded Machine Translation Systems*, Seattle, Washington.

Véronis, J. (ed.) 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora.* Dordrecht: Kluwer.

Voelker, R. 1995. 'Speaking the languages of medicine and culture'. *Journal of the American Medical Association* **273**, 1639–1641.

Wolmuth, P. 1996. 'Removing the barriers'. *Health Visitor* **69**, 93–94.

Woodhead, D. 2000 *The Health and Well-being of Asylum Seekers and Refugees*. London: King's Fund. www.kingsfund.org.uk/asar.PDF

Wynn, R. 1999. Provider–Patient Interaction: A Corpus-Based Study of Doctor–Patient and Student–Patient Interaction. Kristiansand, Norway: Høyskoleforlaget.

Zotti, M. 1999. ‘Public health education for Liberian refugees’. *Nursing and Health Care Perspectives* **20**, 302–306.

Author Index