Managing Multilinguality: Machine Translation and Multilingual Language Technology

Farwell, David
Univ. Politécnica de Catalunya. Centre de Tecnologies i Aplicacions del Llenguatge i la Parla. Jordi Girona Salgado, 1 – 3.
08034 Barcelona
farwell@lsi.upc.edu

BIBLID [0212-7016 (2008), 53: 2; 433-474]

Hizkuntza minoritarioen kasuan, esaterako, euskararen kasuan, Itzulpen Automatikoak honako oinarrizko hiru esparru hauetako bakoitzean izan dezakeen eragina deskribatzen du artikuluak: asimilatzeko itzulpenean, hedatzeko itzulpenean eta interakziozko egoerako itzulpenean. IAren ikerketaren eta garapenaren egungo egoera aurkezten du eta IA sistemen eta IAren errendimenduaren ebaluazioari loturiko gaiak jorratzen ditu.

Giltza-Hitzak: Itzulpen Automatikoa. Hizkuntz aniztasuna. Baliabide mugatuak. Hizkuntza minoritarioak. Euskara.

Este artículo describe el impacto potencial de la Traducción Automática en las lenguas minoritarias, como el Euskera, en cada uno de los tres tipos básicos de la traducción, para la asimilación, para la diseminación y en situaciones interactivas. Presenta el estado actual de la investigación y el desarrollo de la TA y trata sobre temas relacionados con la evaluación de los sistemas de TA y el rendimiento de la TA.

Palabras Clave: Traducción Automática. Multilingualidad. Recursos limitados. Lenguas minoritarias. Euskera.

Cet article décrit l'impact potentiel de la Traduction Automatique sur les langues minoritaires, comme l'euskara, dans chacun des trois domaines de base : la traduction par assimilation, la traduction par dissémination et la traduction en situations interactives. Il présente l'état actuel de la recherche et le développement de la TA et aborde différentes questions liées à l'évaluation des systèmes de TA et le rendement de la TA.

Mots Clé : Traduction Automatique. Multilinguisme. Ressources Limitées. Langues Minoritaires. Euskara.

INTRODUCTION

In a world which each day is growing smaller and smaller, interaction with and interdependence on speakers of other languages is increasingly commonplace. People who as little as two generations ago were unlikely to meet more than a few foreigners during their entire lifetimes now interact on a daily basis with those from other countries and other speech communities. A company in India may be providing telephone customer service or technical support to the clients of businesses in Europe or North America. An Australian athletic shoe company may be acquiring basic materials from Brazil, assembling their products in China and marketing and selling them in France. An emergency room in a large city such as London or Paris may be treating patients from dozens of different countries speaking perhaps over 100 different languages. Multilinguality is now central to the lives of billons of people around the world. It surrounds us. We are faced with language choices when visiting web sites, when using ATMs, when looking at menus, when using urban transportation, and on and on. While shared languages are often found for carrying out ordinary daily intercourse in such situations, they are not always to be found, resulting in at best limited success and quite possibly out and out failure in achieving one's goals.

Multilinguality is also central to Europe. With 25 members and growing, the European Union is now home to some 75 autochthonous languages, 23 of which are official languages of one or another of the member states. Almost half of the official languages have fewer than 10 million speakers and of the additional 53 unofficial languages, 48 have fewer than 2 million speakers. In addition, there are at least another 25 languages spoken by significant immigrant communities including Russian, Chinese, Arabic, Turkish, Ukrainian, Yiddish, Urdu, Bengali, Hindi, Belorussian, Bosnian, Croatian, Macedonian, Berber, Albanian, Armenian and Tatar, just to name a few. The problems faced by government and industry in providing goods and services to people in their native language, in many cases a matter of law, is at the very least daunting.

For business multilinguality presents problems, on the one hand, of internationalization (or globalization) while, on the other, of localization. Products, even for niche markets, are intended to be sold all over the world. Manufacturing, distribution and servicing may need to be integrated on a worldwide basis. Activities may now include:

- gathering information about potential partners or investment opportunities around the world,
- seeking financing and investments from banks or on stock exchanges around the world,
- purchasing component parts and materials from suppliers from around the world,
- preparing and distributing manufacturing specs and servicing manuals to employees around the world,
- advertising in markets around the world,

- selling to customers from around the world,
- providing product information and customer service to clients or users around the world.

For governments and NGOs multilinguality presents problems of on the one hand inclusion and on the other individualization. Basic services need to be provided to everyone in as transparent and effective a manner as possible while at the same time respecting the rights and privacy of each individual. Here activities include, among others:

- circulating public service information to citizens from diverse communities,
- providing health services to a heterogeneous population,
- assisting in tax preparation and collection from a heterogeneous citizenry,
- supporting voter's registration and electoral participation by a culturally diverse electorate,
- electioneering in culturally diverse communities,
- offering legal assistance to members of diverse communities,
- providing formal education to a culturally diverse population,
- explaining the common cultural heritage (history, traditions, customs) of the society as a whole to a heterogeneous population,
- encouraging component communities to maintain their respective cultural heritage within the society as a whole.

For individuals, multilinguality provides barriers on the one hand to effective learning and on the other to full participation. In regard to learning, or more generally information gathering, activities might include:

- following current affairs – reading the news in different languages,
- planning travel – accessing tourism related information in different languages,
- dealing with health issues – consulting on-line medical websites in different languages,
- purchasing goods – reviewing consumer information in different languages,
- entertaining oneself – playing games, watching films, reading literature, etc. in different languages.

In terms of participation, or more generally information exchange, activities may include among others:

- implementing home pages for international access,
- contributing content to YouTube for international access,
- writing book or film reviews (e.g., for Amazon) in different languages,
- writing travel reviews (hotels, restaurants, sights, etc.) in different languages,
- participating in multilingual blogs,
- participating in on-line chat supporting multilingual interaction,

- offering or bidding on items on eBay at an international level,
- sending or receiving e-mail from someone using an unfamiliar language.

Translation, of course, has a crucial role to play in overcoming language barriers and facilitating cross-language communication. The core technology for dealing with this extensive language variation is Machine Translation (MT). It is used (or may be used) for education, science, business, diplomacy and personal communication. While the classic application scenario is the translation house, whether a government organization, a corporate department or an independent private service provider, there are a range of new scenarios in which translation and multilingual technologies will play a central role, including:

- information discovery and recovery (recovering documents or filling out set templates with information relevant to some task - multilingual information retrieval, multilingual information extraction),
- information analysis (sifting through large multilingual text bases for information which supports or refutes hypotheses or formulates predictions of future events – question-answering, multilingual text mining, link detection, hypothesis generation and testing),
- information dissemination (summarization, report generation, and language options for software and web content localization, possibly as an element of the semantic web),
- multimodal interaction (especially chat room and e-mail interaction, multilingual Human-Computer Interaction and intelligent tutoring systems, including language tutors).

In each case, MT or other translation-related technologies and resources must be seamlessly integrated into the work stream to enhance information exchange and increase productivity.

The central goal of this article is to survey research and development in Machine Translation and translation-related technologies although its focus is mainly on MT. At the same time, the perspective of the survey will be from the vantage point of how MT, given the state of the art, might or might not be used to support or promote the use of the languages of smaller speech communities, and in particular Basque, given the current context of rapid globalization.

1. MANAGING MULTILINGUALITY FOR ASSIMILATION

For an assimilation task, that is to say, for information monitoring, gathering and processing with some particular end in mind, the participant is essentially attempting to build a coherent picture of some topic from text or other linguistic source. These may include web-based documents, on-line news services or e-magazines, blogs, e-mail, streamed radio or video, or mobile communication devices. Industry is mainly interested in corporate, financial and economic information as well as possibly in passive Market analysis (say, identifying the opinions of perspective customers). Government might use such sources to compile statistical information or perhaps to monitor potential threats to security. Individuals might use such sources to stay abreast of current events, inform oneself about a health condition, plan travel or decide where to eat or what film to see. For instance, one technology which is currently being developed combines photography, OCR and translation on a PDA platform in order to assist travellers abroad (Gao, et al. 2001). A traveller uses the device to take a picture of a sign and it will then provide a translation on demand. So should the Spanish traveller in the United States wonder if the English language sign reading:

violators will be prosecuted

really means what it appears to mean:

los violadores serán perseguidos/procesados,

he or she needs only to use the PDA for the proper translation:

se procederán infractores.

Such information can actually be very important should you be about to enter a prohibited area, about touch a high tension line or about to drink contaminated water. The same technology is also being applied to reading menus and, a bit further off, to reading tourist brochures.

1.1. Generic information technologies for assimilation

The primary information technologies used to support information assimilation tasks are Question Answering (Webb & Webber 2008), Information Retrieval (Manning, et al. 2008) and Information Extraction (McCallum 2005). Without going into details, the question answering task involves finding the answer to a specific question given a large collection of texts. So if a user asked:

Who is Davey Moore?

the system might return a text snippet akin to:

David S. "Davey" Moore (1 Nov 1933) was an American world-champion boxer who fought professionally 1953-1963 and who died March 25, 1963, as a result of injuries sustained in a match against Sugar Ramos.

Information Retrieval usually takes as a prompt a set of key words and phrases, such as:

Davey Moore, boxer, died from injuries and ideally returns a list of pointers to all and only those texts in the text collection that are about Davey Moore, the boxer, who died from injuries. It is essentially the same process as running a Google search in which case the text collection is the set all documents on the web. Initially, information extraction differs from both question answering and information retrieval in that the prompt is a template of a priori established information needs related to some type of event. So, for instance, one type of event might be "death-from-boxing" and the a priori information needs might include the name of the boxer, the date of birth, the hometown, the weight class, the span of the boxer's career, the bout during which fatal injury occurred, the date of the bout, the opponent, the fatal injury, and the date of death. The text collection is then searched and all and only those texts that describe a "death-from-boxing" event are processed, each bit of relevant information reported in a document being recorded in the corresponding slot in the template. In the end, the user is left with a set of templates describing deaths from boxing.

MT may be embedded in any of these tasks at any of three points. It may be used to translate the query (or prompt) into other languages in order to search for relevant documents in those languages. It may be used to translate documents in other languages into the language of the query or prompt (and presumably the language of the user). In the latter case, the documents may first be translated in order to the process them as if they were in the language of the query or prompt. Alternatively, they may be translated only after being processed in the other language in order to allow the user to inspect the contents of the results of processing. In other words, the system might translate "who is Davey Moore?" into, say, Chinese in order to locate all the Chinese language documents about Davey Moore, translate just those documents back into English and then process them along with all the English language documents about Davey Moore. Or, the system might process the Chinese documents using Chinese language question-answering technology and then translate into English only the resulting snippet.

1.2. Core language technologies

The core natural language technologies underlying the above assimilation tasks may include:

- morphological analysis (e.g., van Halteren, et al. 2001),
- syntactic analysis (e.g., Collins 2003),
- named entity recognition and classification (see Nadeau & Sekine 2007),
- word sense disambiguation (see Aguirre & Edmonds 2007),
- semantic dependency (logical subject, logical object, etc.) labelling (see Màrquez, et al. 2008),
- reference resolution (see Mitkov, et al. 2001),
- event extraction (see Ahn 2006),
- recognizing paraphrase relations (e.g., Barzilay & Lee 2003),

– recognizing textual entailments, (see Giampiccolo, et al. 2008),
– text classification (see Sebastiani 2002),
– text summarization (see Das & Martins 2007),
– speech recognition (see Huang, et al. 2001).

It is assumed the reader is familiar with what the aim of each of these procedures is even if not with the way in which the mechanism operates. Each of these procedures continues to be the object of research though some are considerably more developed and reliable than others. For a general introduction to all aspects of Natural Language Processing, see Jurafsky and Martin (2008). The point here is simply that by mixing and matching such procedures, the assimilative information tasks described earlier can be implemented with a greater or lesser degree of success.

1.3. Core resources

Most of the core language processing technologies rely on a certain set of resources, either in their development or in their operation or both. Generally, the resource fall into one of three types: lexical databases, annotated corpora or simple raw corpus. The most widely used lexical databases are the set of WordNets (e.g., Vossen 1998), mostly derived for other languages from the Princeton WordNet for English (Fellbaum 1998), and the set of FrameNets (e.g., Subirats & Petruck 2003), mainly derived for other languages from the Berkeley FrameNet for English (Ruppenhofer, et al. 2002). In addition, many languages have sharable computational lexicons which may be used to expedite the process of building NLP tools for the language. The major annotated corpora are the different treebanks (e.g., Civit & Martí 2004), modelled on the original Penn Treebank for English (Marcus, et al. 1993) and the different propbanks (e.g., Martí, et al. Forthcoming) modelled on the initial Propbank developed by Palmer, et al. (2005). These resources are especially useful for developing stochastic models of morphological, syntactic and semantic analysis. Finally, there are several large monolingual corpora for most of the world's major languages and several bilingual or multilingual corpora especially for English (e.g., Roukos, et al. 1995), the major European languages (e.g., Koehn 2005) and the major Asian languages (e.g., Sun, et al. 2002).

1.4. Statistical machine translation (SMT)

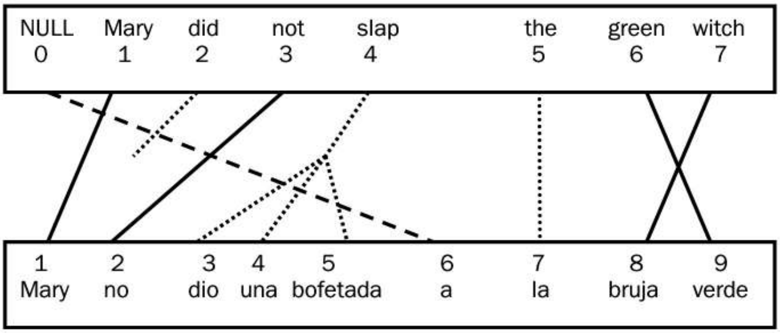
Because translation for information assimilation, or indicative translation, on the one hand, requires large volumes of text to be translated quickly and, on the other, does not require perfectly accurate or fluent translation, especially if the reader has strong background knowledge, the primary approach to this task today is stochastic. Generally, a stochastic approach to the translation process models the likelihood that a particular target language (TL) expression, *t*, is the translation of some specific source language (SL) expression, *s*. The prototypical system consists of three statistical models: an alignment model, a translation model and a target language model for improving TL fluency. See Brown, et al. (1993) for a description of the pioneer SMT effort.

The alignment and translation models are built by looking at very large amounts of parallel texts, millions of words of parallel text if possible. The text is generally broken down into relatively short units (such as sentences).

The alignment model essentially provides statistics to answer questions such as the following: given the word in the third position of the source language expression, what is the likelihood its counterpart is in the first position of the target language expression? What is the likelihood its counterpart is in the second position of the target language expression? What is the likelihood its counterpart is in the third position? And so on, for all combinations of positions. On a slightly more formal level, given that s_i is in the i^{th} position in the SL expression, what is the likely position of its counterpart, t_j, in the TL expression? By simply counting the number of times each case is true in a huge corpus and dividing by the total number of position correspondences, the result is a set of likelihoods for each possible alignment.

In fact, the alignment model itself is composed of three sub-models. The first of these deals with what is referred to as fertility or the number of TL words that correspond to a given SL word. In most cases a given SL word (or phrase) s_n corresponds to a single TL word t_i. But in many cases s_n may correspond to two TL words, t_i, t_j, or even three or more TL words t_i, t_j, t_k,... In Figure 1 below, there are examples, indicated by dotted lines, of items with fertility one, for instance, *the*, with fertility three, e.g., *slapped*, and fertility zero, e.g., *did*. The second component of the alignment model deals with what is referred to as distortion or the possible repositioning of the TL item corresponding relative to the position of a given SL item. Of course it often happens that when s_i occurs in the third position of the SL expression, the corresponding TL item, t_j, occurs in the third position of the translation. But it is also possible that t_j occurs in the first position, or the second position, or the fifth position and so on. The distortion model indicates what the likelihood is for each of these cases. In Figure 1, for instance, *Mary* and its translation equivalent both appear in position 1 as indicated by the solid line, while the counterpart of *not*, in SL position 3, appears in position 2 of the translation and the counterpart of *green*, in position 6, is inverted with the counterpart of *witch*, in position 7, appearing in TL positions 9 and 8 respectively. The final component of the alignment model deals with what is referred to as spurious elements or spurious insertions, that is to say, words that show up in the target language for which there is no SL counterpart. An example of this is shown in Figure 1, indicated by the dashed line, where the use of *a* (to) in Spanish to mark the indirect object of *dar* (give) has no counterpart in English because there is, in fact, no indirect object in the English translation. The statistics are based in this case on creating an artificial empty initial position, 0, in the SL expression which used as the counterpart to all TL words that have no counterpart in the original SL expression.

Figure 1. Types of alignments



The statistics from these three models are then combined to provide a single set of probabilities for the likelihood of any given correspondence between the positions of translation equivalents in source and target expressions.

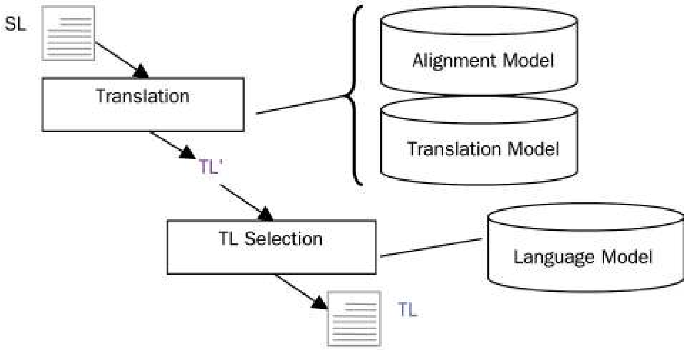
Once the SL and TL texts are aligned, the next step is to induce the translation model. This essentially provides answers to the following questions: given a word s_n in the SL expression (possibly in a specific context), what is the likelihood that its target language equivalent (the aligned counterpart) is t_i? What is the likelihood that it is t_j? What is the likelihood that it is t_k? And so on. The answers can be provided by simply counting the number of occurrences of each type of aligned counterpart of s_i, $t_i \ldots t_{i+n}$, and dividing by the total number of occurrences of s_i in the SL corpus

The alignment model and translation model are then applied together to novel source language expressions. For each SL expression, the combined models suggest various different sequences of words in the target language associating with each sequence a specific likelihood. That is to say, given a sequence of SL words s_i, … s_{i+n}, the combined alignment and translation models are applied to suggest a number of possible translations, t_j, … t_{j+m}, ordered by likelihood. Each will have different words (translations) and different word orders (alignments).

To select the most promising of the suggested translations, a third model, the target language model is applied. Looking only at a large monolingual target language corpus, this model essentially provides statistics to answer the following questions: given two words t_i and t_j in that order, what is the likelihood that the next word in the sequence is t_k? What is the likelihood that the next word is t_l? And so on for all the words of the language. In order to calculate these statistics, every sequence of t_i followed by t_j followed by t_k, for some specific t_k, is counted and the result is divided by the total number of sequences in the corpus beginning with t_i and t_j.

By applying the target language model to each of the translations suggested by the alignment and translation models, the overall likelihoods are modified and the suggested translations are reranked. The highest ranking translation at the end of the process is the selected as the final translation. Figure 2 is a schematic presentation of the translation process.

Figure 2. SMT translation process



On the whole, this approach is very robust, generally intelligible, reasonably accurate but not especially high quality. It improves the larger the parallel corpus there is for training the models because statistical samples for specific expressions are larger and consequently it is less likely a novel combinations of words will be encountered for which there is insufficient statistical information to make suggestions.

1.5. Example systems

Most SMT systems continue as experimental prototypes. An exception to this, however, is the set of systems developed at Language Weaver, a company that was set up in the early 2000's by researchers from the Information Sciences Institute in California with the intent of commercializing SMT (Benjamin, et al. 2002). Initially deployed for Arabic-English translation, it now offers a suite of language pairs including French-English, Italian-English, Spanish-English, German-English, Russian-English, Persian-English, Chinese-English, Korean-English, and Hindi-English. Another interesting development in the direction of commercialization, or at least toward the development of working systems, has been the implementation of Moses, an open source SMT software platform that allows MT researchers and developers to develop their own SMT systems (Koehn, et al. 2007). Developed at the University of Edinburgh and extended under the EU 7th framework project EuroMatrix, it was first released in 2006 and has since been used to implement SMT systems between a large number of the EU's 23 official languages. In Spain, researchers at the Universidad Politécnica de Valencia have developed a statistical MT system based on inducing a finite state transducer from parallel corpus (Casacuberta & Vidal 2004). The procedure is to use statistical alignment of a parallel corpus to produce a set of equivalent expressions from which a stochastic rational grammar is then inferred. The grammar, in turn, is converted into a finite state transducer which is used as an engine for translation.

1.6. Major challenges

Some of the basic challenges facing research and development of stochastic systems include:

- how to approximate theoretically accurate calculations during processing so that the statistics are reliable but at the same time the calculations are possible within constraints of time and memory,
- how to exploit linguistic information for inducing the statistical models,
- how to deal with translation between languages with limited digital resources,
- how to develop large annotated multilingual parallel corpora or comparable corpora,
- how to integrate fully automatic black box MT systems into the overall translation process.

In purely theoretical terms current statistical models are rather reliable. But there are so many possibilities to calculate that at run time even modern CPU and storage capacities are such that it might take hours or even days to actually compute the likelihoods of the possible translations of a given SL text. Thus, one problem facing statistical MT researchers is how to approximate the theoretically valid statistics with models that are at once reliable and computationally practical.

Another set of research issues concerns how to improve the different models on the basis of linguistic information beyond simple word forms, sequences of word forms and word form alignments. Here, even minimal extensions such as to word stems, part-of-speech or morphosyntactic information (case, number, noun-adjective agreement, verb-nominal agreement, etc.) have sometimes lead to improved performance (although not necessarily). For example, there have been experiments with different alignment techniques which focus on sentence "segments" (e.g., Deng, et al. 2004), that is to say, coarse sentence internal constituency such as adverbial, relative or participial clauses. Experimentation has also begun using linguistically motivated morphosyntactic analysis (e.g. POS category with morphological features), shallow syntactic analysis (e.g., constituent chunks), syntactic analysis (e.g. syntactic trees) and even semantic analysis (e.g., semantic dependency graphs). The statistics are modeled using currently available annotated corpora such as the Penn Tree bank or Propbank for English or corpora produced from scratch using currently available NLP analyzers. For instance, Yamuda and Knight (2001) examined a technique for developing a syntactic transfer system using aligned corpora in which at least one of the languages is syntactically annotated.

Beyond this, there has been interest in developing statistical models for languages that may not have a good deal of parallel corpus. For instance, recently experiments have been carried out in which alignment and translation models for a given SL-TL pair have been built by combining alignment and translation models for the SL and a third language, an intermediate language (IL), and alignment and translation models for the IL and the TL where it happens that there is a sufficiently large bilingual parallel SL-IL and IL-TL corpora. Another thrust has been to investigate the possibility of using comparable, rather than parallel, bilingual corpus for training the alignment and translation models (e.g. Munteanu, et al. 2004). Here the SL and TL corpora are generally in a like genre and in a similar domain (e.g., business news) but are not translations of one another.

Finally, statistical MT systems are usually fully automatic black box translators and cannot easily be integrated into the work stream of a particular translator. But the quality, while much improved, is not especially good. Thus such systems are generally used to support document filtering for assimilation tasks such as information analysis, email and chat specifically for texts in languages unknown to a "customer." In this case, the system provides its translation such as it may be and the customer must decide whether the document appears to be worth closer investigation, in which case it is passed to a human translator. In other words, statistical systems are used to replace the translator whose work is rather to focus on especially interesting documents requiring high quality translation or possible on post-editing.

1.7. The case of smaller speech communities

Indicative translation might help to promote the use of languages having smaller speech communities such as Basque. Indicative translation into, say, Basque would encourage the use of the language by those involved in information gathering tasks such as reading company web pages or filtering blogs for how a product is perceived. Indicative translation out of Basque would allow non-Basque speakers access to information expressed in Basque. This might motivate more people to use Basque in their daily activities since they could assume they are reaching a broad audience. On the other hand, it might inhibit expansion of speech group since the need to learn the language would be diminished.

The general problem for languages having smaller speech communities or limited digital resources is that there is a lack of data available for training up statistical systems. For Basque in particular, there are additional complications. As a morphologically rich language it has many more morphological variants for a given lexical item. Where in English you might have *book* and *books* as the variants of the noun *book*, in Basque you have *liburu*, *liburua*, *liburuak*, *liburuan*, and so on up to possibly 68 different forms. Because of this, the size of the corpus must be much larger than for an analytic language such as English because the statistical models depend on seeing multiple examples of each case. In addition, because of the agglutinative nature of Basque, a single word form in Basque often corresponds to a sequence of words in analytic languages such as Spanish or English. The fertility of Basque with respect to say English would be rather high, again making the alignment problem more difficult as well as requiring a much larger training corpus. Yet another complication which Basque presents in the training of statistical models is its relatively free word order with respect to other languages such as English. In other words, there will be a rather high degree of distortion. The alignment model must therefore consider many different permutations requiring a very large training corpus in order to find a sufficiently large statistical sample for each case. Finally, there is the problem of pronominal ellipsis. With respect to English, this mean there will be a fairly high degree of spurious insertions which is also problematical for inducing the alignment and translation models.

So Basque presents serious challenges to building virtually every component statistical model of a stochastic translation system. Nonetheless, there are possible approaches to dealing with some of the issues mentioned.

2. MANAGING MULTILINGUALITY FOR DISSEMINATION

Translation for dissemination, or informative translation, presents a number of problems that can be glossed over by systems used for assimilation. Generally, these are all related to the need for high quality translation. Because of that need, systems designed for dissemination must be able to handle context in terms of shared knowledge and culture and not simply language correspondences.

This is perhaps best exemplified by some less than felicitous translations of advertising slogans that have been documented over the years (although documented, they may nonetheless be heretical). For instance, for one of its earliest US marketing campaigns for its vacuum cleaners, Electrolux used the slogan:

Nothing sucks like an Electrolux

Unfortunately, the verb *sucks* has various meanings and the preferred meaning in this case is *there is nothing quite so undesirable as an Electrolux*. Pepsi Cola entered the Chinese soft drink market with a translation of it very successful US advertising slogan:

Come alive with the Pepsi generation.

Unfortunately, the translation was understood as:

Pepsi Brings Your Ancestors Back From the Grave.

The Gerber baby food company has long used the smiling baby as its icon on the individual serving jars it sells. When it began distributing its products in Africa the results were not positive. It turns out that in Africa, companies routinely label products with pictures of what is inside because many people do not read. Buyers were therefore confused if not upset about what they thought the jars contained. Perdue chicken, an American food processing company, began marketing its product in Mexico with a Spanish version of its slogan:

> *It takes a strong man to make a tender chicken.*

That translation was:

> *Hace falta un hombre potente para hacer un pollo tierno.*

American Airlines, to advertise its new leather business class seats on flights to and from Mexico, used the slogan:

> *Vuela en cuero.*

And Coca-Cola's product name was initially transliterated into Chinese as:

> *Ke-kou-ke-la*

which, it turns out, means *bite the wax tadpole* (or *female horse full of wax*). So they changed it to:

> *ko-kou-ko-le*

meaning *happiness in the mouth*, a definite improvement.

The need for dealing with shared knowledge and culture in order to achieve the quality needed for informative translation is also reflected in the following. Consider the two differing translations into English of *el tercer piso* and *el segundo piso* which appear in the following Spanish fragment from a news article about the Moscow real estate market in the early 1990's.

Source text:

> *. . . los 300 metros cuadrados del* ***tercer*** *piso estaban disponibles pero fueron aquilados . . . , sólo queda el* ***segundo*** *piso . . . .*

Translation 1:

> *. . . the 300 square meters of the* ***third*** *floor were available . . . , but they were rented . . . . All that is left is the* ***second*** *floor . . . .*

Translation 2:

> *. . . the 300 square meters on the* ***fourth*** *floor were available, but they were rented . . . ; only the* ***third*** *floor remains . . . .*

While one translator has rendered these expressions as *the third floor* and *the second floor* respectively, another has rendered them as *the fourth floor* and *the third floor*. Although these two translations are clearly different, they are, in fact, both accurate and they are not necessarily logically inconsistent. The explanation can be found in the differing assumptions the translators have about what the author of the Spanish text believes about the world, and what the audience of the English translation believes about the world.

Essentially, the first translator assumes that the author of the SL text shares the translator's floor naming convention, for lack of a better expression. That is to say, they both think the levels of a building, starting at ground level, are refered to as the *ground floor*, the *first floor*, and so on all the way up. What is more, the first translator assumes that the addressees of the translation also share the translator's floor naming convention. Thus, the first translator refers to the fourth level above ground as *the third floor* and the third level above ground as *the second floor* just as he or she assumes the SL text author did. If these assumptions are correct, then the first translator's translation is accurate and the translator is communicating to the audience of the translation what the author of the SL text intended to communicate to his or her readers.

The second translator, on the other hand, assumes there is an alternative floor naming convention, namely, one for which the levels of a building, starting at ground level, are refered to as *first floor*, the *second floor*, and so on all the way up. In addition, the second translator assumes that either the author of the SL text does not share the translator's floor naming convention or, alternatively, the addressees of the translation do not share the SL text author's floor naming convention. Thus, the second translator refers to the fourth level above ground as *the fourth floor* and the third level above ground as *the third floor*. If either of those sets of assumptions is correct, then the second translator's translation is accurate and he or she is communicating to the audience of the translation what the author of the SL text intended to communicate to his or her readers, even though the translation may differ from the the SL text in terms of its semantics.

A somewhat less mundane example of the effects of culture context in informative translation concerns consumer produce information. In bilingual Canada, with legal requirements for bilingual packaging, the preparation instructions on a package of rice are interestingly different. The English version reads:

> *Add $^1/_2$ cup rice per **$1^1/_2$ cup** boiling water.*
> *Bring to boil.*
> *Simmer for **10 minutes**.*

The French version, on the other hand, (translated back into English) reads:

> *Add $^1/_2$ cup rice per **1 cup** boiling water.*
> *Bring to boil.*
> *Simmer for **8 minutes**.*

Basically, it appears from these two sets of instructions that English speakers generally prefer softer, mushier rice (cooked longer and with more water) whereas French speakers prefer their rice more *al dente*. The "translation" correctly takes this cultural variation into account.

In regard to Business, the primary translation tasks for dissemination are related to localization, the marketing of a product or service offered internationally to a local community having its own special social and cultural characteristics. This often involves language translation. We have seen that it is very important in advertising and for consumer product information. It is also relevant for a wide range of corporate documentation, including service manuals, manufacturing specifications and part lists, for companies having a worldwide manufacturing operation.

For government, translation for dissemination is important for the individualization of its services. In areas with multilingual populations, translation for dissemination plays a role in providing citizens with accessible information about legal rights and obligations, health, transportation, taxes, and so on. It may also play a role in electioneering, voter registration and even as part of the voting process. Beyond this, it may play an important role in education, especially in areas with significant immigrant populations. Additionally, it may serve a double function in maintaining cultural and social norms, on the one hand, of the larger society and, on the other, of the component groups making up that society.

Finally, at the level of the Individual translation for dissemination facilitates participation. Such modern phenomena as personal home pages, MySpace areas or YouTube presentations can be made accessible to an international audience when supported by reliable, high quality translation.

2.1. Generic information technologies for dissemination

Given the primary function of informative translation is to prepare texts (e.g., manuals, brochures, instructions, packaging, specs, ads, etc.) for distribution to clients, customers, users, employees or simply other peers, the core language technology in which translation might be embedded is text generation (Hovy 2000; Pattabhiraman & Cercone 2007). Very broadly, text generators take as input some sort of formal specification of the content to be expressed by the text to be generated. There are three points in this process that MT might be integrated. One is in the generation of the content representation. This might be done by applying the analysis component of a transfer or interlingual MT system to a normal natural language text expressing the desired content. A second point of application is at the level of content representation itself. This might be given directly, or in some modified form, to the generation component of an MT system into the language of interest. The third point of application would be to the output text of the generation system. Here, full MT from the language of the output of the generator into the languages of interest would be used.

2.2. Core language technologies

The basic language processing technologies that are used for text generation are:

– lexical selection (e.g., Habash 2004; Barzilay & Lapata 2005),
– sentence planning (e.g., Stent, et al. 2004; DiMarco, et al. 2006),
– generating referring expressions (e.g., Krahmer, et al. 2003; Gatt 2006),
– surface realization (e.g., White, et al. 2007; Wong & Mooney 2007),
– speech synthesis (e.g., Oh & Rudnicky 2002; Bonafante, et al. 2006).

In addition, there are two approaches to informative MT which have been successfully used albeit under rather special circumstances. The first of these relies on the preparation of documents, generally by human technical writers, using what are referred to as "controlled languages." These are normative forms of a language specifically designed for the preparation of particular document types (engineering specs, service manuals, etc.) for a specific audience (engineers, mechanics, etc.). Controlled languages are special in that they usually have a rather limited or even closed vocabulary consisting for the most part of unambiguous lexical items. They are also restricted syntactically (e.g., no or limited types of recursion) and style (e.g., few pronouns, short sentences, etc.). The advantage of controlled language documents from the perspective of MT is that systems can be tailored specifically to the constraints of the controlled language. While still not perfect, they are sufficiently accurate and high quality so as to reduce the overall time and cost of preparing the foreign language versions of the documents. The approach was originally adopted by Xerox in the 1980's for the translation of user manuals and since by several major international companies.

A second successful approach to translation for dissemination involves "sublanguages", that is, highly limited, naturally occurring versions of ordinary natural languages. They are quite common for domain specific daily news updates such as weather reports, stock market reports, sports summaries, and other brief, highly repetitive text types. Because they are so abbreviated and so focused in terms of domain, the sublanguage itself has a very limited or even closed vocabulary, which is mainly unambiguous, and a simple syntax often consisting of short sentences or even sequence of phrases with little or no recursion. Again, given the restricted nature of these sorts of texts, full linguistic and generally accurate coverage is feasible, making the development of a high quality MT system possible. This approach was first followed in the development of the MÉTÉO for the translation of daily weather reports from weather stations around Canada. A more recent example of the use of sublanguage is the multilingual generation system developed at CoGenTek (Kittredge 1986) which translates stock market reports.

2.3. Rule-based machine translation

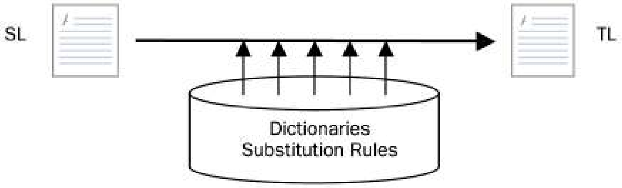
As mentioned above, statistical translation while robust and sufficiently accurate for many assimilation tasks is not especially high quality. An alternative and more traditional approach to developing translation systems is to attempt to model the process as a series of rule applications. Such "rule-based" approaches explicitly encode one or more levels of linguistic knowledge, including lexical, morphological, syntactic, semantic and/or even pragmatic knowledge in the form of formal rules which are then applied in some ordered way to an input text in order to produce some sort of representation of the text which in turn is used to produce an output text. Rules may be lexical or grammatical, the latter often taking the form of phrase structure rules:

A →B C | D_E

either with or without context.

Rule-based systems are generally classed according to their overall system architecture. The three basic types are direct, transfer and interlingual. Direct systems substitute words or sequences of words in the SL input text directly for their corresponding TL translations, moving unit by unit from beginning of the input text, usually a sentence, to the end (see Figure 3).

Figure 3. Direct (substitution) translation

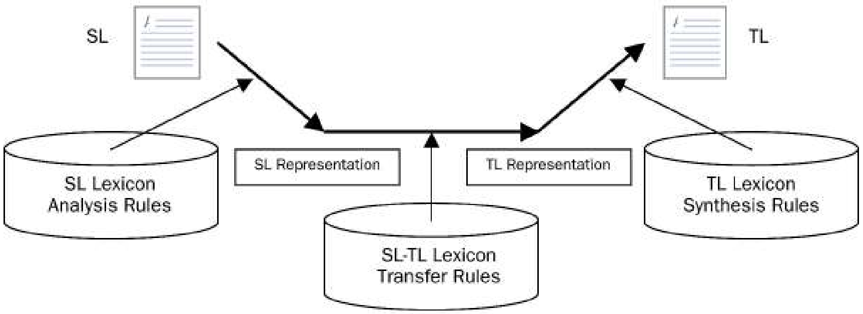


Transfer systems apply a set of analysis rules, generally morphosyntactic but possibly semantic as well, to the input SL text in order to produce a linguistic tree-structure representation. This is followed by the transfer phase during which a set of rules is applied that substitutes SL lexical items by TL lexical items and modifies the tree structure representation where needed.

The final step is to apply a set of text realization rules which inflect the lexical items and insert the requisite function words (prepositions, conjunctions, determiners, etc.) to produce an appropriate and fluent TL text. See Figure 4.

Interlingual systems, akin to transfer systems, apply a set of analysis rules, not simply morphosyntactic but semantic and possibly pragmatic as well, in order to produce a representation of the information the speaker/author intended to express. In the second step, this meaning representation is used as the input the generation component which selects appro-

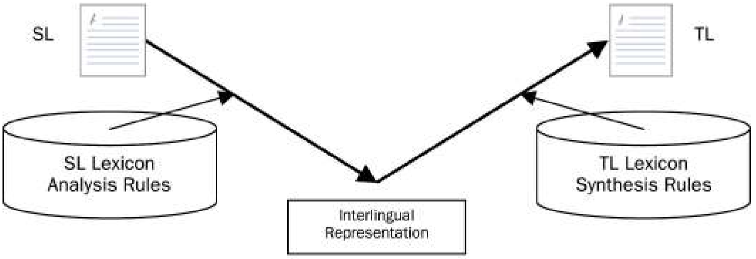
Figure 4. Transfer-based translation



priate TL lexical content, plans the general organization of the TL text (the number of clauses and nominalizations, the embeddings, the co-reference relations and types of referring expressions) and then applies morphosyntactic realization rules, including rules inserting requisite function words, in order to produce an appropriate and fluent TL text. See Figure 5 below.

There are some cases of applying fully automatic rule-based systems to dissemination tasks (e.g., job descriptions) especially if boring, repetitive, closed domain translation is involved. Typically, in these cases, the translation is automatically generated by the MT system and then passed along to a human post-editor, ideally monolingual, who produces the final fluent version. In any case, the key here is that the task involves a domain-limited, repetitive translation task and that the automatic translations are sufficiently high quality to make post-editing more efficient (and less expensive) than human translation.

Figure 5. Interlingual translation



2.4. Example systems

Most commercial systems today are rule-based. The oldest, but still very respected, of these is Systran, the first translation company established in 1968. Initially it was a direct system but over the years has undergone extensive revision and currently most closely approximates a transfer system. It has been developed for some 30 different language pairs involving some 15 different languages and, together, these are being used to translate millions of words of text annually. Around 2000, under the brand name Babel Fish, Systran was adapted for use by the AltaVista internet search engine, which has eventually ended up in the hands of Yahoo!. A second rule-based system of note is the TAUM MÉTÉO, mentioned earlier, which was developed between 1975 and 1977 at the Universitè de Montreal (Chevalier, et al. 1978). It is a transfer system which was used until recently to translate weather bulletins from some 200 weather stations around Canada from English into French. It went on line in 1978 and its success, and it was very successful, providing very high quality translation of 99% of the hundreds of reports it received daily was due to the sublanguage of the genre. Because of the telegraph-like syntax, closed unambiguous vocabulary, limited use of prepositions and complete lack of pronouns, it was possible to provide a complete formal description of the sublanguage of the bulletins. Another, more recent, rule-based MT system is the Kant controlled language system developed over a 6-year period beginning in 1990 at Carnegie Mellon University (Mitamura, et al. 2001). As a controlled language system, its success depends on the careful observation of document preparation guidelines by technical writers within large organizations. Such guidelines generally limit sentence length, vocabulary and grammatical constructions so as to make the documents explicit and easy to read. Kant was deployed at the Caterpillar heavy equipment company in 1997 and has been successfully incorporated into Caterpillar's translation workstream, reducing overall translation costs. As a final mention of a successfully operating rule-based MT system, the SoftLibrary Spanish-Catalan system was developed for *El periodico de Catalunya*, a bilingual daily newspaper published in Barcelona. It is a direct translation system exploiting a bilingual phrasicon of equivalent expressions drawn from a bilingual parallel corpus of news articles. The system was incorporated into the paper's daily edition workstream in 1997 and has been undergoing modifications since. Articles are initially prepared in Spanish and then automatically translated into Catalan. While post-editing is essential, it is sufficiently minimal so as to allow the full translation of the daily two hours before going to press.

2.5. Major challenges

The primary problem of rule-based systems is that they are very costly in terms of the amount of time and expertise needed to build the rule base. In addition, since we do not have complete and comprehensive formal descriptions of any language, texts often fall completely, or in part, outside the purview of the system. In addition, many aspects of language are simply not rule-governed or at least not grammatical. Thus, there is a growing interest in:

- automatic or semi-automatic techniques for building or extending grammars and lexica,
- the development of reusable resources,
- introducing robustness into the analysis and generation phases,
- techniques for improving target language fluency,
- mechanisms for modelling pragmatic phenomena.

To deal with the complications and costs of constructing very large rule bases and lexica, interest has turned to developing automatic or semi-automatic techniques. Not only should this speed up the process of construction, but it should help to reduce the complexity especially reducing ad-hoc rules for very specific situations. An alternative approach to reducing the development effort for rule-based systems is to promote reusable rule-bases and lexica which could be adapted and extended by one group and then returned to a first group in an improved form.

Because there will always be data which falls outside the scope of a system's grammar and lexicon, (e.g., misspelling, new words, ill-formed text, etc.) and there will always be texts for which there are multiple possible translations (e.g., *wooden tables and chairs* as [*wooden* [*tables and chairs*]] or as [[*wooden tables*] *and chairs*]), improving the robustness of systems when faced with unforeseen language and how to improve the ability of systems (or more likely their users) to adapt to new domains and new terminology is crucial. The interest here is in developing mechanisms for producing a best translation in spite of the fact that the system cannot do this solely on the basis of it rules and lexicon.

Why in English we *level scores*, and *buildings* and *charges* but not *records* or *complements* is not especially predictable. It is just how we say things. Because of such conventions of language use there are many decisions rule-based systems must make that are not naturally captured by rules. What is needed here is a mechanism, akin to a style checker, which replaces unnatural collocations with fluent collocations.

Beyond this, there has been a good deal of interest in developing techniques for dealing with a range of semantic and pragmatic phenomena, modelling contexts and reasoning within these contexts to interpret and translate non-literal language, metaphor, metonymy or resolving references.

2.6. The case of smaller speech communities

Informative translation might help to promote the use of languages having smaller speech communities such as Basque. Informative translation into, say, Basque would encourage the use of the language by promoting localization efforts on the part of businesses and efforts to individualise services on the part of governments. This would motivate Basque speakers, non-native and learner as well as native speaker, to use Basque in carrying out their daily activities. Informative translation out of Basque would, if companies and public organisations prepare their documents in Basque, facilitate globalisation and inclusion activities. While this might not promote the use of Basque on the part of non-Basque speakers, it would certainly promote the well-being of the members of the speech community. Again, in the latter case the ultimate effect might be to inhibit expansion of speech group since there would be less need to learn the language.

Rule-based systems, in fact, are precisely the approach that most MT developers have taken to implementing systems for language with smaller speech communities and with limited digital resources. On the one hand, they have been around longer, before the renewed interest in stochastic approaches and before there were extensive bilingual digital resources, not to mention the computational capacity to process with them. Rule-based approaches can be implemented without reference to large digital resources, and they can be extended and modified in a relatively transparent manner. The problem has not so much been in theory as in practice. As mentioned, they are very expensive to develop and they are brittle in the face of language they may not have been programmed to handle. The result has been that translation systems into or out of the languages of smaller speech communities, including languages such as Danish, Norwegian, Finnish, Slovakian, Slovenian, and Irish, have been very limited until rather recently.

With respect to Basque, there has been a fairly active MT community given its size and resources. Among the rule-based approaches, IXA, the Natural Language Processing research group at the University of the Basque Country, has been involved for some time in developing multilingual language tools for localization and are currently participating in two Spanish-government funded MT projects, OpenTrad and OpenMT, to develop open-source MT systems and multilingual tools for various languages of Spain including Basque. See Alegria, et al. 2007, for a description of their Basque-Spanish transfer-based MT system. Another effort is being carried out by InterLan/Geinsa, a company based in Bilbao. It has for some years been developing an interlingual MT system for several languages including Basque, English, Spanish and German.

3. MANAGING MULTILINGUALITY FOR INTERACTIVE SITUATIONS

Translation for multilingual interaction presents problems even more complex than translation for dissemination. Not only is there a need for a context consisting of world and cultural knowledge against which a text is interpreted, a context for the discourse, but there is a need for a more limited but dynamic context which is updated as the interaction proceeds. What is more, unlike the basic discourse, this context, i.e., the context for the speaker's current utterance, must be able to maintain beliefs and reason with beliefs about the world which actually contradict the knowledge represented in the discourse context. It is also necessary to understand people's goals in uttering something and strategies being used to achieve them.

As an example of the effects of utterance context on translation, consider the possible translations of *Feydeau* in the following sentence uttered during a scene from the motion picture, *Jesus of Montreal* (Arcand 1989):

Hein, on va pas jouer une scène de ***Feydeau***

Depending on different views of the background knowledge of the film's audience (that is to say, the discourse context for the addressees of the subtitles) and different views of the utterance context, the possible alternative subtitles include:

> *We are not acting out a scene from* ***Feydeau****.*
> *We are not acting out some scene from* ***a bedroom farce****. / This isn't* ***a bedroom farce****.*
> *There is no danger in our being discovered.*

Under the right circumstances, any of the translation might be most appropriate. How can this be so?

Let's suppose the translator is sitting beside us interpreting the film as it develops, each character in the film acting as an independent agent. Thus, for any given utterance, there are four relevant participants: the actor who speaks, the actor who is addressed, the translator who is observing the interaction, and the audience, the addressees of the translation who ideally should be observing the interaction in the same way as the translator if the translator is successfully performing his or her task.

As background (i.e., that part of the utterance context of which the protagonists, the translator and the film's audience are all aware), the following is a synopsis of what has transpired so far.

A priest at a shrine outside Montreal has been sponsoring a religious drama every summer for 35 years. Since the text has become somewhat outdated, he asks Coulombe, a young actor who has recently returned from an extended sojourn, to modernize the script and to play the part of Jesus. He agrees and immediately sets about looking for collaborators. The priest suggests that Constance, an old friend of Coulombe's, would be a good person for Coulombe to enlist in his endeavor. Among others, Coulombe seeks out Constance and she agrees to work with him. In passing, invites him to stay at her apartment and he agrees.

In a later scene, the scene in which the sentence above is uttered, Coulombe returns earlier than usual to what he assumes is an empty apartment. He starts to make himself comfortable, making some noise in the process. At this point he hears someone moving about in Constance's bedroom and, suddenly, she emerges from within, closing the door behind her. She says, *T'es déjà là, toi?* (Back already?) and then, coughing significantly, says to herself, *Bon...* (Okay...). At this point Coulombe realizes that there is someone else in the bedroom and whispers, *Tu veux que je m'en aille?* (Should I go?). She shakes her head no, laughs nervously, opens the door and says to whomever is inside, *Ben, écoutes, sors* (Come on out), *On va pas jouer une scène de Feydeau*.

It is this last utterance and its subtitle that is at issue here. The translator who provided the subtitles for the film has glossed *On va pas jouer une scène de Feydeau* as *This isn't a bedroom farce*, the second of the options presented above. The question is what are the underlying assumptions that determined the translator's choices in each alternative case.

At the time of Constance's utterance, the protagonists, the translator and the audience have the following beliefs among others (please bear with us).

Coulombe is living in Constance's apartment.
They are close friends and colleagues.
Coulombe has entered the apartment unexpectedly early.
It is still mid-afternoon.
He accidentally makes a loud noise.
Constance emerges from her bedroom dressed in a nightgown and closes the door behind her.
She is somewhat flustered by Coulombe's unexpected presence.
Coulombe believes there is someone else in Constance's bedroom and that he has caught them in a compromising situation.
Coulombe believes that Constance and the other person might prefer some privacy.
Coulombe believes that Constance and the other person might wish to keep the identity of the other secret.
Constance believes that Coulombe believes that she and the other have been caught in a compromising situation.
Constance believes that Coulombe believes that they might prefer some privacy.
Constance wishes to change Coulombe's belief.
Constance tells Coulombe not to leave.
Constance tells the person in the bedroom to come out.

Without going into the details, the analysis of the utterance begins by establishing, on linguistic grounds, that Constance is using *on va pas jouer un scène* to express to the person in the bedroom that she does not wish the current situation (such as Coulombe's discovery of her and the unknown person alone together in her bedroom) to be understood as being a scene from a to-be-determined type of play (i.e., *we're not acting out a scene ...*, *we're not going to act out a scene ...*). The next step is to assign an interpretation to *de Feydeau*. Again, on linguistic grounds, coupled in this case with knowledge of the world, we establish that Constance is using *de Feydeau* to refer to the type of situation that might be described in a play by the 19th century French playright Feydeau who wrote bedroom farces (i.e., *a scene from a bedroom farce*). To arrive at this interpretation, it must be the case that:

Constance believes Feydeau is a playwright and that Feydeau wrote bedroom farces.

Constance believes the person in her bedroom believes Feydeau is a playwright and that Feydeau wrote bedroom farces.

The interpretation is completed by confirming that the situation under discussion (i.e., Constance and someone alone together in her bedroom) is indeed one that Feydeau may have written about. This becomes especially plausible when it turns out that the man in Constance's room is a priest, in fact, the very same priest who hired Coulombe to update the play.

Having arrived at an interpretation, the translator now needs to provide an equivalent expression for an English speaking audience. To express that some current situation is not of some type, he/she selects the expression *This is not a scene ...* or *This isn't a scene ...* or some such English equivalent. As for a situation typical of a bedroom farce of the sort that Feydeau might write about (i.e., two people getting caught in a compromising position by a significant other), the translator checks his/her beliefs about the background knowledge of the audience. This leads to the first case of variation in translation stemming from variations in the utterance context, namely, those based on variations in the translator's assumptions about the addressee of the translation, the non-French speaking audience of the film. If it is assumed that the addressee of the translation would not typically know that Feydeau is a playwright or that Feydeau wrote bedroom farces, quite possible for anyone unfamiliar with French culture or with the theater, then reference to Feydeau will fail to have the intended effect and some alternative expression must be chosen, e.g., *a bedroom farce*. In the event that the translator assumes that the film's audience has the same beliefs about Feydeau as the speaker (Constance) and the addressee (the unknown person in the bedroom), he or she would most likely take advantage of those beliefs to provide a translation that more closely approximated the source language utterance in form and content, relying on the addressee of the translation to use those beliefs appropriately to interpret Constance's utterance. That is, if the translator assumes the addressees of the translation believe that Feydeau is a French playwright who wrote bedroom farces, then he or she would most likely have glossed the utterance as *This is not a scene from Feydeau*.

A second case of translation variation due to differences in the utterance context are similarly based on the translator's assumptions about the what addressees of the translation know about the world. In the first case, the translator needed the necessary knowledge about Feydeau to work out what Constance meant in uttering *On va pas jouer une scène de Feydeau*, namely, that the kind of situation they found themselves in could be, but was in fact not, typical of a play by Feydeau. Thus, the translator was able to provide any of the translations mentioned above such as *This isn't a scene from Feydeau* or *This isn't a bedroom farce*. But beyond this intended meaning, the translator also needs the necessary knowledge to figure out why Constance said what she said, namely, to inform the unknown person in the bedroom that Coulombe would not be scandalized by their liaison and that the unknown person in the bedroom could safely show himself. It is equally necessary that the audience of the translation have the requisite knowledge after reading the translation to work out Constance's motive. While rather unlikely, if the translator believed that the audience of the translation did not know what a bedroom farse is, then he or she would not be able to use that expression either. In such an event, the translator might simply have explicitly spelled out the motivation. This may be accomplished by glossing *On va pas jouer une scène de Feydeau* as *There's nothing to worry about*.

The implications of this discussion is that the assumptions of the translator and the way in which the translator reasons from them underlie the eventual form of the translation. Those assumptions and the associated reasoning therefore determine translation quality. It should also be clear that there is a wide range of potentially appropriate translations for a given interaction since variations may arise from differences in participants' knowledge and that every participant (translator, author, reader and audience) has a different and incomplete knowledge of the individuals, objects, situations and events referred to in a communicative interaction.

An approach to MT that takes such pragmatic factors into account offers the only direct assault on the issues raised by Bar-Hillel as early as 1959 (Bar-Hillel, 1960). It is not simply that MT systems need knowledge, they need to be able to create complex structures of assumptions and to be able to reason within those structures in order to arrive at appropriate interpretations and translations of not just the information content but of the goals and strategies of the participants in the interaction as well. This, in spite of possibly having incomplete or possibly inconsistent knowledge.

The general task involved in translation in interactive situations is information exchange, that is to say, on the one hand, assimilating information from elsewhere and, on the other, disseminating information elsewhere in response. For business, interactive translation may play an important role in such activities as banking, commercial transactions, automated telephone receptionists or customer service or in making hotel, auto rental, air travel or theater or concert ticket reservations. Trivial examples include the use of multilingual menu systems for automatic bank tellers or for fast track airline check-in. There are a number of activities in the public sector as well that require, or at least would benefit from, interactive translation. These include several major governmental activities such as census data gathering, tax collection, voting, medical visits, emergency room triage or other such activities as might require simultaneous interpretation. Tax preparation services may be offered as a matter of law in the official languages of municipalities, provinces or countries which are multilingual, and even where there are no legal obligations, interactive translation in support of tax preparation has been shown to increase the total amount of taxes collected. Finally, for individuals, activities related to social networking would potentially benefit from interactive translation as well. For instance, reading and posting music, book or product reviews on sights such as Amazon or hotel, restaurant or other travel-related reviews on sites such as Travelocity or Yahoo! would reach a broader audience if accessible in multiple languages. Similarly, contributing additional text to Wikipedia entries in a range of different languages would potentially reduce the total human effort dedicated to their translation and possibly add consistency to entries across languages. Other activities that would benefit greatly from cross linguistic interaction include offering or bidding on items on e-Bay, participating in blogs or chat rooms or exchanging e-mail.

3.1. Generic information technologies for interactive situations

The major applications for interactive dialogue systems include database interfaces, expert systems, tutoring systems, and games. Expert systems are quite common now especially for supporting medical diagnosis and planning treatment as well as research. They generally consist of a knowledge base, say associations of symptoms with deceases, a reasoning engine capable of making inference on the basis of the knowledge in the knowledge base, and a natural language interface to facilitate the system's use by allowing users to interact with it without learning some special query language and by freeing up the user's hands so they can be involved in some other activity at the same time. Using interactive translation in second language learning has been controversial for some time but for those who find it useful, fully automatic MT could conceivably be (and has been) incorporated into on-line reading, writing and translation exercises, including more recently the use of multilingual chat rooms and e-mail correspondence. Whether the system provides high quality translations or merely hints at the content of the source text, it might be used to assist in understanding or producing texts in the language being acquired as well as to provide materials which need to be edited using knowledge of the language to be acquired.

3.2. Core language technologies

The core language processing technologies involved in interactive situations include:

– discourse analysis (e.g., Hobbs 1985; Litman & Passonneau 1995; Carlson, et al. 2002),
– discourse planning (e.g., Hovy 1993; Young & Moore 1994),
– discourse generation (e.g., Bateman, et al. 2001; Soricut & Marcu 2006),
– context modelling (e.g., Ballim & Wilks 1991).

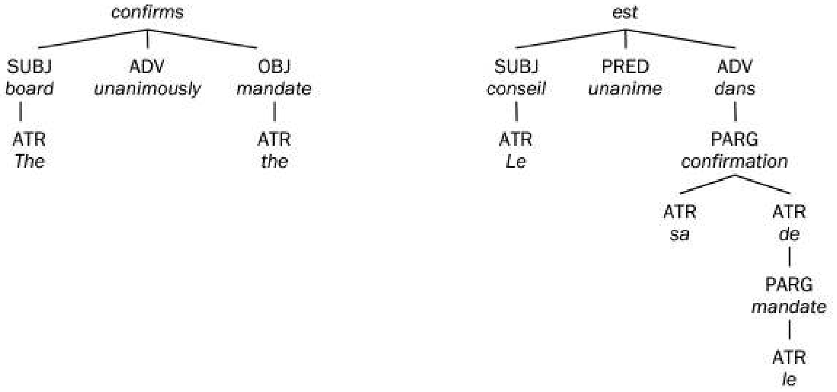
Discourse analysis focuses on assigning a range of grammatical, semantic, pragmatic or rhetorical relations between segments of a discourse, whether in the form of continuous text or in the form of spoken or written interactions (such as e-mail correspondence). Relations vary but core semantic relations concern the relative time and place of the events described and the various sorts of causal relationships they might have with respect to one another. One set of pragmatic relations, for instance, has to do with the types of contribution in an interactive exchange such as asserting, informing, requesting, information seeking, etc. Rhetorical relations are more textual in nature and relate to the structure of a coherent presentation of information. For instance, where one sentence might be viewed as asserting something new, the next might be seen as an elaboration or a consequence or a background condition or a justification or evidence for that assertion. Discourse planning and generation are concerned with first organizing information for presentation in a coherent manner and second with expressing the organized information through the use of language. It determines how to structure information in fairly specific terms, identifying what has been said already and what has already been referred to so as to select appropriate referring expressions, such a pronominals, and avoiding redundancies and repetitions. Context modelling concerns the organization of the background information against which language is interpreted or formulated. It certainly relates to our knowledge of the world, but it also concerns our attitudes toward information being expressed and our assumptions about the knowledge and attitudes of others involved in the interaction (speaker, addressees, those referred to by the text or during the course of conversation). Needless to say, research in this area is incipient and the systems that have been developed are for the most part highly experimental.

3.3. Example-based machine translation

Although perfectly adequate for informative translation (as well as indicative), an approach to MT which has actually been used to develop experimental systems embedded in an interactive task is example-based MT. The prototypical example-based system, like statistical translation, is also corpus based, but it approaches the corpus with different assumptions and different goals. The basic idea is that there is already a lot of high quality human translation available, so why not use it for novel translation. If I have already translated an expression such as *prima por coste de la vida* as *cost-of-living increase*, why not simply use that translation whenever I run across *prima por coste de la vida* again? Assuming I can access and adapt such correspondences quickly, recycling translations can save time while at the same time maintaining greater consistency across translations.

Before an example-based system can be implemented, a necessary preliminary is to prepare a database of example translations, an example base. First, a large bilingual or multilingual parallel corpus must be assembled. Each of the monolingual corpora in the corpus is then analyzed morphosyntactically. For example, the English sentence, *The board unanimously confirms the mandate,* and its French counterpart, *Le conseil est unanime dans sa confirmation de le mandate*, might be analyzed as in Figure 6.

Figure 6. Monolingual morphosyntactic analysis



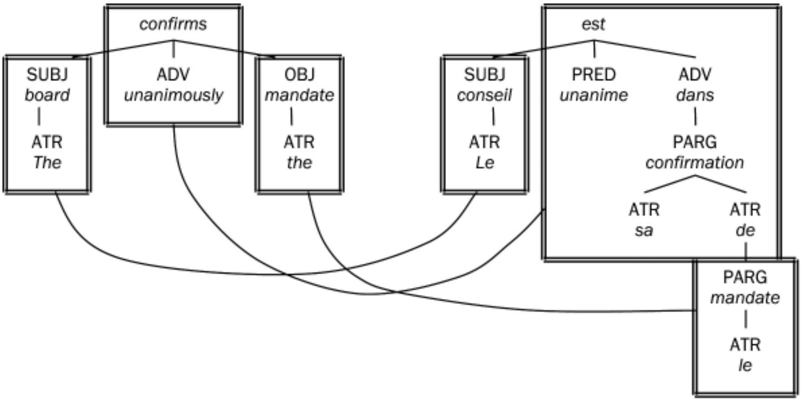
The next step is to align the corpus at the constituent level. For instance, the example English and French sentences would align as in Figure 7.

Figure 7. Multilingual corpus alignment

the board	*le conseil*
the	*le*
unanimously confirms	*être unanime dans sa confirmation*
unanimously	*unanime*
the mandate	*le mandate*
the	*le*

The aligned corpus is then used to identify equivalent translation units. Essentially these are the aligned constituent-level counterparts with preference being given to longer units. The results of the process are shown in Figure 8.

Figure 8. Identification of translation equivalents



Finally, translation units are grouped together on the basis of the similarity of the source language forms. In large corpus, many units will be repeated or very similar. Repeated translation equivalents are filtered. Similar units may be generalized by collapsing example by removing morphological variation or by replacing smaller, less consistent constituents in larger units with variables. But this must be done carefully since the strength of the example-based approach is in capturing subtle translation variations that are quite possibly related to shifts in tense or plurality or collocation. An example of final results, that is, an example base, is provided in Figure 9.

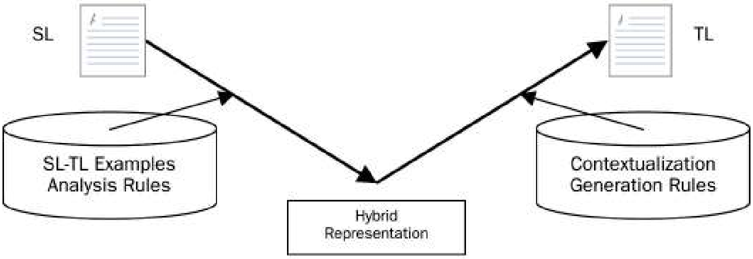
Figure 9. Grouping examples

have a direct effect on	*ont une influence directe á*
have a direct effect on	*interessent directement*
had a direct effect on	*ont eu une répercussion directe sur*
has had a marked effect on	*a largement influence*
had a positive effect on	*s'est avérée positive dans*
had a highly negative effect on	*en auraient été gravement affectés*
will have a decisive effect on	*influencera de façon déterminante*
etc.	etc.

Examples are often simple constituents and generally capture collocational relationships rather nicely. For instance, although the equivalent of *license* in *fishing license* is *permiso* or *licencia* (*de pescar*), the equivalent of *license* in *driver's license* is *carnet* or *permiso* (*de conducir*). Because they are so specific and there are so many of them, this sort of collocational convention is rather difficult to implement in a rule based system. Examples may also be made up of sequences of constituents as in *level a building*, whose equivalent would be *arrasar un edificio*. Compare this with the equivalent of *level the score* (i.e., *igualar el marcador*) or *level charges against* (*hacer acusaciones en contra*). Examples might even be full sentence such as *in for a penny, in for a pound* (i.e., *de perdidos, al río*).

Once the example base has been prepared, example-based translation is carried out in two phases. First, the input SL text is matched, segment by segment, against the example templates on the SL side of the example base and, if a match is found, the corresponding TL template is made available for generation. If no match is found, or if there is untranslated material corresponding to a template variable, then the text may be translated using a conventional rule-based system or some other technique. In some sense, the equivalents in the example base are similar to the equivalents recorded in a translation memory except that many may be so basic that any translator would think them too obvious to warrant recording and others may be so "literal" (e.g. it might include *level a building*, *level a barn*, *level a skyscraper*, and so on) that a translator would undoubtedly collapse them into a single generalized case (such as *level a building*). In the second phase of EBMT, the corresponding TL example templates are "spliced" together to form the translation. Splicing is essentially the process of filling in any template-internal variable elements and adjusting other elements for contextual dependencies between translation segments. See Figure 10 for a diagram of the process.

Figure 10. Example-based translation



This approach was initially developed in the 1980's by the Kyoto University MT research group working on a grammar-based approach to Japanese-English translation (Nagao 1984). It has three principle advantages. First, it allows for the treatment of discontinuous constituents such as *figure out* in English which often appears with intervening material as in *figure the answer out*. Second, it allows the translation system to deal with idiosyncratic collocational phenomena (which all of the examples above reflect). Finally, it offers the translation system increased potential to generate natural, fluent, even colloquial target language text. More recently, it also provides an additional advantage. It can be incorporated more easily into a rule-based MT system (unlike the combination of rule-based and statistics-based systems.

As example-based approaches try to increase the use of the examples during processing (as opposed to applying a general grammar-based MT analyzer/generator), they appear to be slowly converging with statistical approaches (which conversely appear to be moving from a focus on string-level substitution to a focus on constituent-level analysis and substitution). More recently, interest in the example-based MT approach has focused on trying to skip the construction of an example base and instead attempting (attempting) to use parallel corpus directly as a source of examples during the translation process. For this exercise to work, corpus alignment is extremely important, as it is for statistical approaches, although there is perhaps more concern for establishing a constituent-level alignment as opposed to a word- or string-level alignment (e.g., Owczarzak, et al. 2006). In addition, there is a good deal interest in improving the matching process between source text and the source corpus of the parallel aligned corpus (Brown, et al. 2003).

3.4. Example systems

Since first being developed in Japan, EBMT has regularly received greater interest on the part of the Japanese research community. For instance, researchers at ATR in Osaka have been following the EBMT approach for two decades, most recently having developed an EBMT system which enhances a traditional syntactic transfer approach using statistical techniques (Imamura, et al. 2004). One problem with traditional EBMT has been example selection. The idea here is to generate multiple possible translations using standard EBMT and then apply a statistical translation model and target language model to select the best among the possibilities. An alternative approach to English-Japanese MT is the MSR-MT EBMT system developed at Microsoft Research (Brokett, et al. 2002) which learns structured phrase-level example translations from bilingual corpus using an abstract logical representation. Each side of a bilingual parallel corpus is parsed in order to produce a common logical representation of the content, which essentially neutralizes surface lexical and semantic differences between the languages. The example translation units are then assumed to be the expressions in the different languages corresponding to a given element of the common representation. A third example of EBMT is MaTrEx (Armstrong, et al. 2006), a hybrid EBMT system developed at Dublin City University which incorporates certain SMT techniques during example-base construction. To build the example base, a bilingual parallel corpus is first word aligned. Separately, the SL and TL monolingual sub-corpora are each segmented into constituents and the parallel corpus is then aligned at the constituent level. The corpus, the word aligned corpus and the constituent aligned corpus are then all used as potential sources of examples during translation.

3.5. Major challenges

There has been extensive interest in developing example-base MT recently as an alternative data-driven approach to modelling the translation process, in part because it allows for a more transparent integration of standard linguistic knowledge. Still, central issues remain to be resolved including:

- developing techniques for finding and generalizing appropriate examples for the example base,
- developing techniques for selecting the most relevant example during the translation process (i.e., for matching the SL input against the example base),
- developing techniques for integrating examples given the context of their application,
- developing techniques for dealing with input, possibly corresponding to example internal variables, for which no appropriate examples are found.

In corpus, as can be imagined from looking at the sample example base in Figure 9, there are many potential expressions such as *have a direct effect on* and *had a direct effect on* which are so similar as to potentially warrant generalization. But their translations, *ont une influence directe á* and *ont eu une répercussion directe sur*, while similar, are not quite so closely related. The research issue here is to develop techniques which correctly decide when two such examples can be collapsed into one or whether or not expressions such as *influence á* and *repercussion sur* are substitutable. Beyond this, there is a good chance that during the actual translation process there are several possible matches on the SL side of the example base. Suppose, for example, that the input expression is *is going to have a direct effect on*. The question is which of the examples in the example base is best or at least a possible match. Another problem is that when an example is selected it must be integrated with the surrounding context. If *ont une influence directe á* is selected, for instance, and the subject is *la loi* (the law), then *ont* must be modified from plural to singular agreement, i.e., *a une influence directe á*. What are needed are techniques for identifying the contextual dependencies of the examples and adapting the examples appropriately. Finally, putting aside for a moment that a word or phrase table might be treated as part of the example base, a significant part of the input to an EBMT system will not find any match at all in the example base. The research problem in this case is to develop techniques that will provide appropriate target language equivalents of such material which can at the same time be seamlessly integrated into the translation process.

3.6. The case of minority languages

Interactive translation for languages having smaller speech communities such as Basque should have the positive effect of allowing speakers to use the language for a wide range of daily activities including shopping, banking, making travel arrangements, accessing customer service, doctors office visits, negotiating governmental bureaucracies and so on even where those products and services are offered in other languages. It might also encourage the use of the language by native speakers while participating in activities related to social networking, participating in blogs or chat rooms and in exchanging e-mail. Finally, interactive translation for Basque might be used as a component of a Basque language learning system or more simply to support interactive language learning software. It is again unclear however whether the ultimate effect might not be to inhibit expansion of speech group since there would be less motivation on the part of non-speakers to learn the language.

Be that as it may, as mentioned above, languages with limited resources present a problem for statistical MT because there is insufficient data for training the statistical models. This is especially true for Basque given its rich inflectional morphology, relatively free constituent order and propensity for ellipsis. Rule-based systems, on the other hand, are costly and time consuming to develop and tend to be brittle in the face of unexpected input. Example-based MT, however, has no a priori need for large bilingual parallel corpora since the example base may always be built by hand or, more interestingly, by automatic or semi-automatic techniques applied to bilingual resources (dictionaries, term banks, translation memory, etc.), limited parallel corpus and/or large monolingual corpora. Example-based MT also generally presupposes morphological analysis and generation as well as some constituent level analysis and generation. This makes it more amenable to translation into and out of languages with limited digital resources where often existing translation systems are rule-based.

As for Basque EBMT, there have been recent efforts on the part of the research group at IXA to develop Basque-English EBMT (Stroppa, et al. 2006) and early results have been encouraging. Alternatively, research at Deusto has focused on developing translation memories to support human translation (Abaitua, et al. 2001). In many respects similar to EBMT, the approach relies on "bitexts," or parallel text segments, which translators select when faced with analogous SL text. Specifically, researchers have developed techniques for automatically extracting "bitexts," from bilingual corpus, which has required developing interesting constituent-level alignment routines for text segments that possibly extend beyond the sentence (Casillas, et al. 2000).

4. HYBRID SYSTEMS

Beyond research activities related to SMT, RBMT and EBMT per se, there has been significant interest in developing hybrid systems that integrate components of the different approaches so as to maximize the advantages that each has to offer. We have mentioned a few examples above in regard to EBMT. More generally these early efforts are looking at:

- substituting one or another component of rule-based systems (analysis, transfer, generation) with a stochastic counterparts,
- developing stochastically generated lexicons for rule-based systems,
- developing stochastic pre-processor or post-processor to improve rule-based throughput,
- developing statistical rule application techniques,
- providing rule-based pre-processing or post-processing for statistical MT systems,
- developing examples based on statistically aligned constituents as opposed to strings.

For instance, Post and Gildea (2008) have looked at supplementing a statistical target language model for SMT with parsers, possibly rule-based. Costa-jussà, et al. (2007) have looked at rule-based string reordering during preprocessing (so as to closer approximate TL constituent order) to improve alignment as well as translation model application for SMT systems. Research in the area of EBMT is looking at how to use aligned parallel corpus directly as an example base (Brown 2008). In addition, Tinsley and others (2007) are investigating technique for shallow annotation to improve example generalization and selection.

Research on Basque MT has also begun to focus on developing hybrid MT systems. As mentioned above, researchers at IXA, for instance, are involved in the OpenMT project, which aims to develop open-source hybrid models of MT combining RBMT, SMT and EBMT. Early outcomes have included a system which translates an input using three different MT engines, one RBMT, one EBMT and one SMT, and then selects the best result based on a handful of heuristic criteria (Alegria, et al. 2008). A second approach to combining frameworks has been to apply a statistical post-editor to the output of an RBMT system that has been trained on a parallel corpus of the RBMT system's actual output and the hand edited version of that output (Díaz de Ilarraza, et al. 2008).

5. EVALUATION

For as long as there has been research in MT, there has been interest in MT evaluation. Indeed, there has probably been more published in the field on MT evaluation than on MT itself. Nonetheless, with the rebirth in interest in statistical approaches in the early 1990's, there also arose a need to develop an evaluation methodology that could compare system performance on comparable tasks, typically, the translation of texts in a common genre (e.g., news articles) and of a similar length. Initially, the methodology proposed (White & O'Connell 1994) was an outgrowth of prior approaches to the problem which relied on human assessments of the fidelity (preservation of the information content expressed by the SL text), comprehensibility (the understandability of the information conveyed by the TL translation) and fluency (the readability of the TL translation). Especially from the perspective of SMT developers, such evaluations were expensive, time consuming, and required bilingual expertise.

As a result, in the late 1990's an automatic evaluation technique was developed at IBM which, while not especially useful as a diagnostic, was shown to correlate with human judgments of relative quality. BLEU (Papineni, et al. 2002), as the methodology is referred to, is a statistical metric which provides a score between 0 and 1 based on the number and length of text segments in an output translation which match text segments of one or more (human generated) reference translations. It is widely used at this point and helps MT developers by indicating whether a more recent version of their system performs better, on a par with, or worse than a prior version as well as telling them how the performance of their system compares with the performance of others over a common test set. But BLEU has it draw backs not the least of which is the fact that the test sets have to be developed generally by human translators. In addition, it is not very insightful. It does not recognize categories of errors nor the strengths and weaknesses in some broad sense of different systems.

At the same time, other evaluation methodologies have been proposed and there has been at least one effort, FEMTI (Framework for Machine Translation in ISLE - King, et al. 2003), to systematically analyze the objectives of an evaluation and to suggest a range of metrics based on those objectives. From an MT developer's perspective, for instance, evaluation should provide diagnostic information about the different components of the MT system as well as the overall quality of system throughput, generally in terms of the parameters mentioned above, i.e., fidelity, comprehensibility and fluency. From the perspective of an information manager who is considering using MT to provide or assist the company's translation service, evaluation should generally focus on the ease of use as well as quality. On a systems level, this might include an evaluation of the facilities for preparing the SL text for translation, for adding or modifying lexicons, grammars or translation models, and for revising the TL translations. On an operational level, this might include an evaluation of the quantity and quality of human intervention needed, the training or experience of those operating the MT system, and, of course, the overall economic impact of using the system. From a translator's perspective, evaluations should consider the amount of pre-editing and/or post-editing required, the overall savings in time and effort that might accrue, and the impact on the translator's regular work routine. Finally, from the perspective of the consumer of the translation, the evaluation should indicate the usability of the translation and the time and cost of obtaining it.

As for other approaches to evaluating the overall quality of the output TL translations, methods for assessing the fidelity of the translation include comparing the quantity of information expressed by the TL translation with respect to that expressed by the SL text. This approach, of course, requires a bilingual human evaluator who is trained at least somewhat in what counts as "information". Alternatively, evaluators can compare the ability of people to perform some task after having read the SL text as opposed to the TL translation. Here, monolingual subjects with no special training can be used. As for comprehensibility, one approach is to simply ask (untrained monolingual) readers of the output TL translation to indicate how intelligible it is, generally on a scale of 1 to 5, with 5 being fully intelligible and 1 being completely incomprehensible. Others have used certain classic techniques from educational psychology for measuring comprehensibility including Flesch scale analysis or Cloze tasks. For evaluating fluency, one technique has been to ask (untrained monolingual) readers to indicate how easy it is read the text, generally on a scale of 1 to 5, with 5 being easily readable and 1 being completely unreadable. Another approach has been to measure reading times and then follow up with a comprehension test. The reading time is taken as an indicator of fluency while the comprehension test is essentially to make sure the text was understood.

Evaluation of linguistic quality has taken either of two general approaches. One, which looks at what is referred to as "edit distance," focuses mainly of the quantity of linguistic errors. In this case, the actions taken by an editor in correcting an output TL translation are counted. Those actions include adding or deleting words, substituting one word for another or transposing one word or phrase with another. The more actions taken, the lower the linguistic quality of the TL translation. Alternatively, or in addition, a qualitative evaluation of the text may be carried out using error analysis. Here, errors are classed into types (phonological, punctuation, lexical or terminological, morphosyntactic, or stylistic – unnecessary repetition of words or ideas, translation tropes, awkward expressions, etc.). Such evaluations are especially valuable for rule-based and example-based MT system development although they can be useful for SMT system development as well.

6. CONCLUSION

We have presented three basic types of translation, for assimilation, for dissemination and for interactive situations. For each we have looked at a range of activities for which that sort of translation is or could be useful, the related information technologies in which MT might be embedded, the supporting NLP procedures, a particularly relevant approach to MT design, i.e., SMT, RBMT or EBMT, and discussed the special problems of languages with more restricted digital resources and in particular Basque. In addition, we have presented a range of evaluation methodologies which have been used to assess the quality of translation and the utility of MT within a working environment. We have seen that, while the challenges of dealing with a multilingual environment are many, varied and often extremely difficult, MT and ML technologies are steadily improving. Today they can be applied to:

- many open domain translation tasks for assimilation,
- certain cases of closed domain translation tasks for dissemination,
- or, in any case, simply used to facilitate the translation task for humans.

Finally, as has long been the case, many procedures and techniques developed for MT can be applied a range of multilingual tasks not only for translators but for ordinary users as well. Issues of interest here include:

- the application of multilingual lexicons and terms banks to information discovery tasks (IR, IE, multilingual text mining, etc.),
- the development of multilingual named entity recognition, classification and translation (proper names, dates and other temporal expressions, alphanumeric expressions, acronyms, etc.),
- automatic techniques for extending or constructing translation memories especially in new domains.

As for the relevance of MT to promoting multilingualism or the use of languages having smaller speech communities, the situation is less clear. There are two broad situations where this occurs. First, in areas where many people speak the same set of languages, MT might be usefully applied in support of language acquisition for any monolinguals or non-native speakers within the community. The more obvious situation is in areas where different groups speak different languages. Here, MT could provide translation to facilitate daily interactions. But in either case, automatic translation could actually inhibit multilingualism by allowing people to navigate foreign languages and cultures without ever having to confront them or learn the language.

REFERENCES

ABAITUA, Joseba; DÍAZ, Josuka; GÓMEZ, Josu; JACOB, Inés; OCINA, Koldo. "XTRA-Bi: extracción automática de entidades bitextuales". XVII Congreso de la SEPLN, Jaén. *Procesamiento de Lenguaje Natural* 27. 2001; p. 305-306.

AGUIRRE, Eneko; EDMONDS, Philip (eds.). *Word Sense Disambiguation: Algorithms and Applications*. Berlin: Springer, 2007.

AHN, David. "The stages of event extraction". In: *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Stroudsburg, PA: Association for Computational Linguistics, 2006; p. 1-9.

ALEGRIA, I.; DÍAZ DE ILARRAZA, A.; LABAKA, G.; LERSUNDI, M.; MAYOR, A.; SARASOLA, K. "Transfer-based MT from Spanish into Basque: reusability, standardization and open source". In: *Proceeding of CICLING*. Berlin: Springer, 2007.

ALEGRIA, I.; CASILLAS, A.; DÍAZ DE ILARRAZA, A.; IGARTUA, J.; LABAKA, G.; LERSUNDI, M.; MAYOR, A.; SARASOLA, K. "Spanish-to-Basque MultiEngine Machine Translation for a Restricted Domain". In: *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas.* 2008.

ARMSTRONG, S.; FLANAGAN, M.; GRAHAM, Y.; GROVES, D.; MELLEBEEK, B.; MORRISSEY, S.; STROPPA, N.; WAY, A. MaTrEx: "Machine Translation Using Examples". In: *Proceeding of the TC-STAR OpenLab on Speech Translation*. Trento, Italy, 2006.

BALLIM, Afzal; WILKS, Yorick. *Artificial Believers: the Ascription of Belief*. Hillsdale, NJ: Lawrence Erlbaum, 1991.

BAR-HILLEL, Y. "The present status of automatic translation of languages". In: *Advances in Computers 1*. 1960; p. 91-163.

BARZILAY, Regina; LAPATA, Mirella. "Collective Content Selection for Concept-To-Text Generation". In: *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing* (HLT-EMNLP). 2005.

BARZILAY, R.; LEE, L. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In: *Proceedings of the HLT-NAACL*. Edmonton, Canada, 2003; p. 16-23.

BATEMAN, John; KAMPS, Thomas; KLEINZ, Jörg; REICHENBERGER, Klaus. 2001. Constructive text, diagram and layout generation for information presentation: the DArt$_{bio}$ system. *Computational Linguistics.* 2001; 27 (3): p. 409-449.

BENJAMIN, Bryce; KNIGHT, Kevin; and MARCU, Daniel. *Translation by the Numbers: Language Weaver*. Berlin: Springer, 2002.

BONAFONTE, A.; ESCUDERO, D.; RIERA, M. 2006. "La conversión de texto en habla". In: LLISTERRI, J.- MACHUCA, M. J. (eds.) *Los sistemas de diálogo*. Manuals de la Universitat Autònoma de Barcelona, *Lingüística*, 45. 2006; p. 177-208.

BROCKETT, Chris; AIKAWA, Takako; AUE, Anthony; MENEZES, Arul; QUIRK, Chris; SUZUKI, Hisami. English-Japanese Example-Based Machine Translation Using Abstract Semantic Representations. *Proceedings of the 19th International Conference on Computational Linguistics.* Taipei, Taiwan, 2002.

BROWN, P. F.; DELLA PIETRA, S.A.; DELLA PIETRA, V.J.; MERCER, R.L. "The mathematics of statistical machine translation: parameter estimation". *Computational Linguistics.* 1993; 19 (2): p. 263-311.

BROWN, Ralf. "Exploiting Document-Level Context for Data-Driven Machine Translation". In: *Proceedings of the 8th Conference of the Association for Machine translation of the Americas.* 2008.

BROWN, R.; HUTCHINSON, R.; BENNETT, P.; CARBONELL, J.; JANSEN, P. 2003. "Reducing Boundary Friction Using Translation-Fragment Overlap". In: *Proceedings of the Ninth Machine Translation Summit*, New Orleans, LA, 2003; p. 24-31.

CARLSON, Lynn; MARCU, Daniel; OKUROWSKI, Mary Ellen. "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory". In: Jan van Kuppevelt and Ronnie Smith (eds.) *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers. 2002.

CASACUBERTA, Francisco; VIDAL, Enrique. "Machine Translation with Inferred Stochastic Finite-State Transducers". *Computational Linguistics.* 2004; 30 (2): p. 205-225.

CASILLAS, Arantza; ABAITUA, Joseba; MARTÍNEZ, Raquel. "Recycling annotated parallel corpora for bilingual document composition". In: John S. White (ed.), *Envisioning Machine Translation in the Information Future, Proceeding of the 4th Conference of the Association for Machine Translation in the Americas.* 2000.

CHEVALIER, M.; DANSEREAU, J.; POULIN, G. *TAUM-METEO: Description du Système*. Université de Montréal, 1978.

CIVIT, Montserrat; MARTÍ, Mª Antonia. "Building Cast3LB: A Spanish Treebank". *Language and Computation.* 2004; 2 (4): p. 549-574.

COLLINS, Michael. "Head-Driven Statistical Models for Natural Language Parsing". *Computational Linguistics.* 2003; 29 (4):, p. 589-637.

R. COSTA-JUSSÀ, Marta; CREGO, Josep M.; LAMBERT, Patrik; KHALILOV, Maxim; R. FONOLLOSA, José A.; MARIÑO, José B.; BANCH, Rafeal E. 2007. "Ngram-based statistical machine translation enhanced with multiple weighted reordering hypotheses". In: *Proceedings of the Second Workshop on Statistical Machine Translation.* 2007; p. 167-170.

DAS, Dipanjan; MARTINS, Andre F.T. "A Survey on Automatic Text Summarization". *Literature survey for Language and Statistics II*. Carnegie Mellon University, November 2007.

DENG, Y.; KUMAR, S.; BYRNE, W. "Bitext Chunk Alignment for Statistical Machine Translation". *CSLP Tech Report.* Johns Hopkins University, 2004.

DÍAZ DE ILARRAZA, Arantza; LABAKA, Gorka; SARASOLA, Kepa. "Statistical Post-Editing: A Valuable Method in Domain Adaptation of RBMT Systems for Less-Resourced Languages". In: *Proceeding of the Workshop on Mixing Approaches to Machine Translation.* San Sebastián, 2008.

DIMARCO, C., COWAN, D.; BRAY, P.; COVVEY, D.; DI CICCIO, V.; HOVY, E.H.; LIPA, J.; MULHOLLAND, D. A. "Physician's Authoring Tool for Generation of Personalized Health Education in Reconstructive Surgery". In: *Proceedings of the AAAI Spring Symposium on Argumentation for Consumers of Healthcare.* Stanford University, CA, 2006.

FELLBAUM, Christiane (ed.). *WordNet: An Electronic Lexical Database*. Cambridge: MA:MIT Press. 1998.

GAO, Jiang; YANG, Jie; ZHANG, Ying; WAIBEL, Alex. "Text Detection and Translation from Natural Scenes". *Technical Report CMU-CS-01-139, Computer Science Department*. Carnegie Mellon University, June 2001.

GATT, A. “Structuring knowledge for reference generation: A clustering algorithm”. In: *Proceedings of the 11th Meeting of the European Association for Computational Linguistics*. 2006.

GIAMPICCOLO, Danilo; MAGNINI, Bernardo; DAGAN, Ido; DOLAN, Bill. 2008. “The Third PASCAL Recognizing Textual Entailment Challenge”. In: *Proceedings of the Workshop on Textual Entailment and Paraphrasing.* 2008; p. 1-9.

HABASH, Nizar. “The Use of a Structural N-gram Language Model in Generation-Heavy Hybrid Machine Translation”. In: *Proceedings of the Third International Conference of Natural Language Generation.* 2004.

VAN HALTEREN, Hans; ZAVREL,Jakub; DAELEMANS, Walter. “Improving Accuracy in NLP through Combination of Machine Learning Systems”. In: *Computational Linguistics.* 2001; 27 (2): p. 199-229.

HOBBS, Jerry R. “On the Coherence and Structure of Discourse”. In: *Report No.CSLI-85-37.* Stanford University: Center for the Study of Language and Information, 1985.

HOVY, E.H. “Automated discourse generation using discourse structure relations”. In: *Artificial Intelligence.* 1993; 63: p. 341-385.

HOVY, E.H. “Language Generation”. Entry for *Encyclopedia of Cognitive Science*. London: McMillan, 2000.

HUANG, X.; ACERO, A.; HON, X. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall. 2001.

IMAMURA, K.; OKUMA, H.; WATANABE, T.; SUMITA, E. **“**Example-based machine translation based on syntactic transfer with statistical models”. In: *Proceedings of the 20th International Conference on Computational Linguistics.* 2004.

JURAFSKY, Daniel; MARTIN, James H. *Speech and Language Processing* (2nd Edition). Prentice Hall. 2008.

KING, M.; POPESCU-BELIS, A.; HOVY, E. “FEMTI: “Creating and using a framework for MT evaluation”. In: *Proceeding of the 5th Conference of the Association for Machine Translation in the Americas.* 2003; p. 224-231.

KITTREDGE, R. 1986. “Variation and Homogeneity of Sublanguages”. In: R. Kittredge and J. Lehrberger (eds.), *Sublanguages: Studies of Language in Restricted Semantic Domains.* New York: de Gruyter, 1986; p. 107-137.

KOEHN, Philipp. “Europarl: A Parallel Corpus for Statistical Machine Translation”. *Proceedings of the 10th Machine Translation Summit.* Phuket, Thailand, 2005; p. 79-86.

KOEHN, P.; HOANG, H.; BIRCH, A.; CALLISON-BURCH, C.; FEDERICO, M.; BERTOLDI, N.; COWAN, B.; SHEN, W.; MORAN, C.; ZENS, R.; DYER, C.; BOJAR, O.; CONSTANTIN, A.; HERBST, E. Moses: “Open source toolkit for statistical machine translation”. In: *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics.* 2007; demonstration session.

KRAHMER, E.; VAN ERK, S.; VERLEG, A. “Graph-based generation of referring expressions”. *Computational Linguistics.* 2003; 29 (1): p. 53-72.

LITMAN, Diane; PASSONNEAU, Rebecca. 1995. “Combining multiple knowledge sources for discourse segmentation”. In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics.* Morristown, NJ: Association for Computational Linguistics. 1995; p. 108-115.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. 2008. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

MARCUS, Mitchell P.; SANTORINI, Beatrice; MARCINKIEWICZ, Mary Ann. “Building a Large Annotated Corpus of English: The Penn Treebank”. In: *Computational Linguistics.* 1993; 19 (2): p. 313-330.

MÀRQUEZ, Lluís; CARRERAS, Xavier; LITKOWSKI, Kenneth C.; STEVENSON, Suzanne. 2008. “Semantic Role Labelling: An Introduction to the Special Issue”. In: *Computational Linguistics.* 2008; 34 (2): p. 145-159.

MARTÍ, M.A.; TAULÉ, M.; BERTRAN, M.; MÀRQUEZ, L. AnCora: Multilingual and Multilevel Annotated Corpora. Forthcoming.

MCCALLUM, Andrew. “Information Extraction: Distilling Structured Data from Unstructured Text”. *ACM Queue.* November 2005; p. 49-57.

MITAMURA, T.; NYBERG, E.; BAKER, K.; SVOBODA, D.; TORREJON, E.; DUGGAN, M. “The KANTOO MT System: Controlled Language Checker and Knowledge Maintenance Tool”. In: *Proceedings of NAACL.* 2001.

MITKOV, Ruslan; LAPPIN, Shalom; BOGURAEV, “Branimir. Introduction to the Special issue on Computational Anaphora Resolution”. *Computational Linguistics.* 2001: 27 (4): p. 473-477.

MUNTEANU, D.; Fraser, A.; D. Marcu, D. 2004. “Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora”. In: *Proceedings of HLT/NAACL.* 2004.

NADEAU, David; SEKINE, Satoshi. “A survey of named entity recognition and classification”. *Lingvisticæ Investigationes.* 2007: 30 (1): p. 3-26.

NAGAO, M. “A framework of a mechanical translation between Japanese and English by analogy principle”. In: A. Elithorn and R. Banerji (eds.), *Artificial and human intelligence.* Amsterdam: North-Holland, 1984.

NAGAO, M. *Machine translation: how far can it go?* Oxford: Oxford University Press, 1989.

OH, Alice; RUDNICKY, Alex. 2002. “Stochastic natural language generation for spoken dialog systems”. In: *Computer Speech and Language.* 2002; 16 (3): p. 387-407.

OWCZARZAK, K.; MELLEBEEK, B.; GROVES, D.; VAN GENABITH, J.; WAY, A. “Wrapper Syntax for Example-based Machine Translation”. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas.* 2006; p. 148-155.

PALMER, Martha; GILDEA; KINGSBURY, Paul. 2005. “The Proposition Bank: An Annotated Corpus of Semantic Roles”. *Computational Linguistics.* 2005; 31 (1), p. 71-106.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.J. BLEU: “A method for automatic evaluation of machine translation”. In: *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics.* 2002; p. 311-318.

PATTABHIRAMAN, T.; CERCONE, N. “Introduction to the special issue on natural language generation”. In: *Computational Intelligence.* 2007; 7 (4): p. 199-206.

POST, Matt; GILDEA, Daniel. “Parsers as language models for statistical machine translation”. In: *Proceedings of the 8th Conference of the Association for Machine Translation of the Americas.* 2008.

ROUKOS, Salim; GRAFF, David; MELAMED, Dan. *Hansard French/English Corpus*. Philadelphia, PA: Linguistic Data Consortium, 1995.

RUPPENHOFER, Josef; BAKER, Colin; FILLMORE, Charles J. "The FrameNet Database and Software Tools". In: Anna Braasch and Claus Povlsen (eds.), *Proceedings of the Tenth Euralex International Congress.* 2002; Vol. I: p. 371-375.

SEBASTIANI, Fabrizio. 2002. "Machine Learning in Automated Text Categorization". In: *ACM Computing Surveys.* 2002; 34 (1): p. 1-47.

SORICUT, R., MARCU, D. "Discourse generation using utility-trained coherence models". In: *Proceedings of the COLING/ACL.* 2006; p. 803-810. Association for Computational Linguistics.

STENT, A.; PRASAD, R.; WALKER, M. "Trainable sentence planning for complex information presentations in spoken dialog systems". In: *Proceedings of the 42nd Annual meeting of the Association for Computational Linguistics.* 2004.

STROPPA, N.; GROVES, D.; WAY, A.; SARASOLA, K. "Example-Based Machine Translation of the Basque Language". In: *Proceedings of 7th conf. of the Association for Machine Translation in the Americas.* 2006.

SUN, Le; XUE, Song; QU, Weimin; WANG, Xiaofeng; SUN, Yufang. "Constructing of a large-scale Chinese-English parallel corpus". In: *Proceedings of the 3rd workshop on Asian language resources and international standardization.* 2002; p. 1-8. Association for Computational Linguistics.

SUBIRATS, Carlos; PETRUCK, Miriam R. L. "Surprise: Spanish FrameNet!". In E. Hajicova, A. Kotesovcova and J. Mirovsky (eds.). In: *Proceedings of the Seventeenth International Congress of Linguistics* (CIL-17). Prague: Matfyzpress, 2003.

VOSSEN, P. (ed.). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers, 1998.

WEBB, N.; WEBBER, B. *Natural language Engineering, Special issue on interactive question answering: Introduction*. Cambridge University Press, 2008.

WHITE, J.S.; O'Connell, T.A. "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches". In: *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas.* 1994.

WHITE, M.; RAJKUMAR, R.; MARTIN, S. "Towards broad coverage surface realization with CCG". In: *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation.* 2007.

YAMADA, K.; KNIGHT, K. "A syntax-based Statistical Translation Model". In: *Proceedings of the 39th Annual meeting of the Association for Computational Linguistics;* 2001; p. 523-530.

YOUNG, R.M.; MOORE., J.D. DPOCL: "A Principled Approach to Discourse Planning". In: *Proceedings of the 7th International Workshop on Natural Language Generation.* 1994; p. 13-20.

WONG, Yuk Wah; MOONEY, Raymond. "Generation by inverting a semantic parser that uses statistical machine translation". In: *Proceedings of NAACL-HLT.* 2007; p. 172-179.