Transfer review

Information Extraction technology in Machine Translation: IE-guided architecture for MT

School of Modern Languages and Cultures
Centre for translation studies

Bogdan Babych

Supervisors:

Prof. Tony Hartley, Prof. Yorick Wilks

Introduction

Since 1960-s Machine Translation (MT) technology has been trying to achieve high translation quality, comparable to the quality of human translations. The goal of "fully automatic high quality MT" is not yet achieved, but substantial progress has been made in terms of involving extensive linguistic resources (lexicons and grammars) and developing efficient MT architectures, which deal with real texts. Recently developed data-driven approaches to MT, such as Example-Based or Statistical Machine Translation, techniques of using monolingual corpora in combination with large electronic dictionaries – have made a major breakthrough in acquiring linguistic resources for MT automatically. However, just using more extensive resources cannot help overcome certain fundamental limits on improving MT quality. It has been recognised that models and architectures for translation need to be improved and the results from other fields of the Natural Language Processing need to be integrated into MT.

Recent developments in the area of Information Extraction have placed this technology in some respects ahead of MT: it focuses on limited and clearly defined tasks, such as Named Entity recognition, co-reference resolution, word-sense disambiguation, filling pre-defined templates, etc., and also has well-defined evaluation procedures. The performance achieved by IE systems on these limited tasks is higher than the performance of MT systems. In these cases, integrating IE analysis into MT causes substantial improvement in translation quality.

These results also suggest that MT research and development need a similar clearly defined and focused strategies, which allows the developers to concentrate on the most important aspects of translation quality, i.e., on resolving issues which cause the most serious factual errors. IE technology addresses the issue of retrieving facts from unrestricted text, therefore it could be possible to use its focused "fact tracking" capabilities for MT technology – e.g., for dealing with factual errors and omissions.

IE technology also inspires thinking of MT in terms of communicative goals. On the one hand, translation of a source text can be guided not only by stored translation equivalents, but also by ways it changes recipient's knowledge about some events (or ways information is extracted into IE templates). On the other hand, the quality of a target text is acceptable only if it can change recipient's knowledge in a similar way as the target text (or if similar information for IE templates can be extracted from it). It will be reasonable to suggest that this interpretation of IE technology can provide a theoretical ground for a systematic development of a communicatively guided MT architecture, which could complement the current equivalent-based approaches and "pragmatically" motivate translation transformations (similarly to transformations in human translations, which are motivated by communicative goals).

In our project we will investigate the requirements of the communicatively guided model for MT and the ways how IE technology can be extended to meet these requirements.

1. Problem statement

An increasing number of studies show that synergies between different areas of NLP have a great potential for improving existing technologies. Current diversity of approaches to NLP makes it necessary to ensure that the developments in one area are not overlooked in other areas [Chesterman, 1998].

There are several examples of successful integration of different NLP technologies or interpreting certain technologies as metaphors for others, e.g., Information Retrieval, Question Answering and Natural Language Understanding were treated in the framework developed for Statistical Machine Translation, or successful application of MT technology in Cross-Language Information Retrieval [Berger and Lafferty, 1999], [Berger et al., 2000], [Macharey, et al., 2001], [Gachot, et al., 1988].

In this PhD project I will investigate possible synergies between Information Extraction and Machine Translation. There are different possible ways to explore the challenge of bringing together IE and MT technologies.

On the one hand MT can be a useful extension to monolingual IE systems, allowing templates to be extracted from multilingual sources. In [Wilks, 1997: 7-8] two directions for such extensions are suggested: MT can either be used for translating texts before submitting them to an IE system, or it can be used for translating filled templates. Within this approach there are many issues that still need to be investigated. e.g., how does the quality of MT influence the performance of an IE system, how do statistical properties of MT affect IE modules that use statistical methods? We addressed some of these issues in our paper [Babych, Hartley, Atwell, 2003]. However these results can also be used for MT evaluation: IE-related measures are reliable indications of some aspects of MT quality, when applied to the target text. In our project we will investigate possible IE metrics for evaluating MT and explore the possibility of integrating these metrics into MT architecture, e.g., for evaluating multiple translation choices.

Different aspects of IE technology can be useful for research and development of MT systems. Two main arguments for using IE methods in MT can be found in the literature.

The first argument is that many problems for MT systems, such as monolingual morphological, lexical and syntactic disambiguation, co-reference resolution, NE recognition and classification, etc., are not unique to translation, they are present in many other NLP applications, including IE [Hutchins, 2003: 505]. The performance of several IE systems on some of these tasks has been systematically evaluated (and as a result significantly improved) at DARPA competitions (MUC-6, MUC-7). Algorithms such as PoS tagging, parsing and semantic representation have also been developed and tested in the IE paradigm. The key components of IE systems are evaluated separately, so technological bottlenecks could be identified for complex IE tasks (such as scenario template filling or summary generation). Research and development of MT can benefit from this focused systematic approach to evaluation, where different aspects of the resulting quality are clearly localised.

The second (perhaps more theoretical) argument for integrating IE techniques into MT is related to the "perfectionist" approaches, such as Knowledge Based MT [Carbonell and Tomita, 1987], [Nirenburg and Carbonell, 1978]. These studies argue for the need to go beyond "core" linguistic areas, and to incorporate discourse pragmatics, and some AI techniques [Somers, 1992: 192]. In our project we will show that existing IE technology is capable of addressing some of these issues. The results of our experiment with improving MT quality with automatic NE recognition [Babych and Hartley, 2003] suggest that IE has a potential for improving some weak points of MT systems. We will generalise this approach and explore ways how existing IE can be extended to meet the "discourse pragmatics" and AI needs of MT systems. Interpreting IE tasks in the framework of discourse pragmatics (more specifically – speech act framework) could become a theoretical ground for such generalisation.

2. The proposed approach

2.1 Rationale

The PhD project will explore the design of an IE-guided architecture for MT, where IE techniques and extended in order to meet the demands of MT. Specifications will be developed for:

- MT-oriented annotation of the source text for an IE system
- IE-oriented evaluation measures for the target text
- integration of the IE annotation and IE evaluation metrics into MT systems

A prototype MT-oriented IE system will be developed for annotating source texts for MT in the subject domain of football match reports, by extending and reusing existing IE resources (the GATE and/ or MUMIS IE systems). The improvement of MT quality achieved with this annotation will be evaluated.

A theoretical framework for IE-guided MT architecture will be explored which should predict possible improvement of MT and should systematically account for possible synergies between IE and MT technologies.

The IE-guided architecture for MT addresses a set of problems related to *communicative excessiveness* of linguistic representations in the process of translation. We suggest that communicative excessiveness is a property of a text where some morphosyntactic and semantic representations do not have illocutionary effects (are not intended to be recognised as extralinguistic communicative intentions of the speaker), and consequently cannot receive "communicative" interpretation, i.e., they are not used to perform any speech act. We also suggest that the communicative excessiveness is dynamic: the "latent" and "active" layers of semantic representations can be different in different situations of the text's use, and are difficult to predict automatically for a new text, so it often becomes a source of ambiguity.

An essential aspect of human "understanding" is the ability to distinguish the "active", (communicatively relevant) and "latent" (communicatively excessive) information layers. This difference becomes evident in the process of translation: translating the "active" layer is obligatory, and translation equivalents used for this must perform the same "speech act", and have the same (or at least a very similar) perlocutionary effect (i.e., extralinguistic consequences of the text, e.g., changing recipient's knowledge about events) in the target culture.

On the contrary, current MT architectures are not designed to dynamically distinguish communicatively relevant and communicatively excessive layers of information. Instead, they use “static” heuristics hard-wired in their databases of translation equivalents. As a result, the choice which information is translated depends on a constant set of available translation equivalents, rather than on a changing communicative goal. The heuristics work fine in many cases, e.g., semantically uninterpretable morphosyntactic features, such as morphological gender, rarely need to be translated. But in some cases this information appears to be communicatively relevant and needs to be translated (see examples in *section 2.3*). On the other hand, the information which is normally assumed to be communicatively relevant, (e.g., lexical information) in some cases is communicatively excessive and can be dispensed with, in order to preserve communicatively “active” layers in translation. And finally, for some kinds of information it is difficult to predict whether or not it will be communicatively relevant in a given text (or in a different situation of interpreting the same text). E.g., information conveyed by word order in Slavonic languages sometimes is relevant for translation into English, sometimes is not, and may require very different types of translation equivalents under different circumstances (articles, word order, lexical means, etc.) so “static” heuristics for such information crash in nearly all cases.

Mistranslations of Named Entities in MT (e.g., ‘Bill Fisher’ translated as “to send a bill to a fisher”) can be also ascribed to the problem of communicative excessiveness. (In the example above: perhaps, Bill Fisher's ancestors were fishers by profession – this kind of information is expressed in the source text, but it is "excessive" in most cases, it is not intended to be recognised as the speaker's communicative intention. However, available sets of translation equivalents make MT systems translating this "latent" layer. The "active" information layer, e.g., that "Bill Fisher" is a noun phrase used as a name of a person, is expressed simultaneously with the "latent" layer. Both layers together cannot be translated into a target language). For the "equivalent-based" MT architectures the chance of choosing a wrong layer for translation is very high, since there is no explicit annotation to dynamically distinguish communicative excessiveness / communicative relevance of linguistic representations in the source text.

But if the "active" information layers are properly identified, annotated and mapped to the target language using appropriate translation equivalents, then the illocutionary effects of the source and target texts are similar or very close in the source and target cultures, and the translation is "adequate". Failure to identify or convey the "active" layers often results in disruption of adequacy (factual errors or omissions) or fluency (stylistic errors). (There is a chance that such translation is fluent or at least adequate, but there is no guarantee that translated "latent" information will generate units of some other "active" information layer, which have similar illocutionary effect in the target culture). On the contrary, translation of the "latent" information layer is not obligatory. Human translators often choose to preserve some of this information in translation, but they do not have to. Translators are free to choose any means in the target language that satisfy the requirements of "*fluency*" of the target text, but which express different information than is expressed in the source text. This freedom causes a certain degree of legitimate variation in human translations. The variation can be measured by lexical N-gram distance scores, such as BLEU [Papineni et al., 2001] or IE-oriented lexical distance scores [Babych, Hartley, Atwell, 2003], and tends to be more or less constant for independently produced human translations of the same text. In our project we will try to interpret these distance scores in terms of communicative excessiveness and develop IE-oriented MT evaluation measures based on this interpretation.

So, it is not possible to predict which information in the text is "active", and which is "latent" without a pragmatic analysis, that should evaluate the illocutionary effect of the text in a specific situation of its use – obviously it is a difficult task for an MT system, – and it often becomes the basis of different types of ambiguities that lead to factual errors and omissions or disruptions of fluency in MT.

However, IE technology is capable of addressing this problem. Many existing IE systems process news reports, and (to some extent) model "assertive" speech acts. The purpose of an IE system is to single out the information which really changes the state of knowledge about the described events (performs an "assertive" speech act having a perlocutionary effect) and to ignore any kind of "redundant" information not aimed at this task. This distinction can be used to guide MT systems to concentrate on really "important" ("active") pieces of information, and not to get mislead by communicatively redundant representations – the strategy can be to ignore them and to use some other "fluency ensuring" target language structures instead. Existing IE systems can be extended to supply MT-oriented annotation of the "active" information layer, as well as to deal with other types of speech acts, so the range of texts for IE-guided MT architecture can be also widened.

The need to go beyond "core" linguistic knowledge (morphological, syntactic and semantic) and to integrate discourse pragmatics and AI techniques into MT is recognised by MT researchers. However, the argument for this integration was mainly focused on certain advanced MT tasks such as anaphora resolution [Nirenburg and Carbonell 1987], [Somers, 2003, p. 518]. In our project we will show that adequate pragmatic model is essential for the core aspects of MT – identifying necessary levels of translation equivalence for particular contexts. Further we will explore ways of developing such pragmatic model on the basis of IE technology that is extended with additional annotation levels (such as an intermediate translation from a natural language to a formal query language, e.g., SQL) in order to address such demands of MT.

We will try to show that factual errors and omissions of the state-of-the-art MT systems can be attributed to the systems' attempts to translate "latent" information and the failure to identify the "active" pieces of information. Different IE tasks can be used to deal with this problem, including NE recognition, scenario template filling from the source text, evaluating the target text with IE-related evaluation measures for MT. We will specify how the annotation produced by the IE systems should be extended to meet some specific requirements of MT in this direction and how IE-related evaluation procedures can be integrated into the MT architecture (e.g., in order to choose the best translation candidate or to force re-analysis of low scoring translations).

2.2. IE-guided architecture for MT

I. In the first stage of our project, the specifications for IE-guided MT architecture will be developed. These specifications will be based on the following series of experiments:

2.2.1. MT-oriented annotation of the source text

Improving the analysis of the source text using the information taken or deduced from IE templates (NEs, strings and annotation used for template element filling, scenario template filling, and co-reference annotation). This will allow specifying necessary MT-oriented annotation for IE systems, which will define formal descriptions of units that need to be recognised, as well as translation strategies for these units. In this direction the following experiments will be carried out:

1.1. Improving MT with IE from the source text

This experiment is aimed at identifying and correcting factual errors and omissions in MT using NE, scenario template, and co-reference annotation of the source text. It will extend earlier experiments on improving MT with NE recognition, and address the question if the analysis of the source text (ST) can be improved with different IE techniques, what are the limits of this improvement and how IE annotation can be extended to meet the requirements of MT. It will be investigated if the information in IE scenario templates or other types of IE annotation can improve MT quality.

The relation between different types of IE annotation in the source text and different translation strategies will be examined. Improvement of MT quality is expected for:

- syntactic segmentation with IE-identified template elements;
- co-reference resolutions for pronouns;
- NEs by flexible selection of translation strategies.

The results of the experiment will be used:

- to specify requirements for the MT-oriented template structure in the subject domain of match reports;
- to specify IE annotation of the ST, which can be mapped into the required translation strategies more accurately.

The results will also provide further evidence for developing a theoretical basis of MT improvement with IE techniques – e.g., whether the model of communicative excessiveness can be efficient. GATE-1 data for the DARPA competition and the MUMIS system [Saggion et al., 2002] will be used.

1.2. Improving MT with a (target-language) IE from a parallel text

The experiment will use a parallel corpus of football match reports, aligned on the word level. MUMIS IE system will provide annotation for NEs, template elements and co-reference relations in the text originally written in English. The information about the word level alignment will be used to generate similar IE annotation in parallel non-English texts. Appropriate IE annotation will be mapped to the non-English part of the corpus: NEs and other template elements will be identified and co-reference relations will be projected from the corresponding aligned structures in English texts. A multilingual IE annotation will be created as a result. This annotation will be used to guide the analysis of the non-English text to improve the quality of MT (similarly to the experiment 1.1.). In addition, the aligned corpus will provide information about preferable translation strategies for IE-annotated items (NEs, template elements and co-referring expressions), which will suggest ways of extending IE annotation to meet the needs of MT systems. The results will specify the requirements for MT-oriented IE systems.

2.2.2. IE-oriented evaluation measures for the target text

Evaluating the target texts with IE-based measures that will model how well the target text fits the IE requirements. These experiments will:
- define criteria how useful the target text is for IE purposes;
- model statistical parameters of communicative excessiveness and communicative relevance of the information in the target text (TT).

This could possibly lead to creation of new IE-based evaluation measures for MT. The following experiments will be carried out in this direction:

2.1. Comparing statistical N-gram models of the MT corpus and the human translation corpus

This experiment extends our research on comparing unigram models in the human translation and MT corpora. These models will allow carrying out a kind of a "cloze test" for MT output: predicting N-th word in MT based on N–1 previous words and the N-gram model for the human translation corpus. It will be investigated whether the results of such "cloze test" correspond to other MT evaluation measures, and can be used for ranking different translation candidates (e.g., produced by different MT systems for the same segment) – e.g., human judgements about MT quality. The goal of this experiment is developing an evaluation measure for MT that does not require human translation to produce the scores.

2.2. Checking integrity of lexical chains in MT

Lexical chains are sequences of repeated notional lexical items (and also – sequences of co-referring expressions) in text. They can be used for summary generation and other IE-related tasks. This experiment will check if the structure of lexical chains is preserved in human translations (or if it is subject to legitimate variation), and how lexical chains behave in MT output. The experiment will show if the integrity of lexical chains can become an IE-oriented evaluation measure for MT and human translation. If this is the case, this evaluation measure will have an advantage of not requiring human translations of the ST to produce evaluation scores for the MT output.

2.3. Measuring the degree of variation in human translation and MT

The BLEU scores and other lexical distance measures can be used to characterise the degree of legitimate variation in human translation. Translation is controlled by the structure of the ST, but a certain degree of variation is to be expected, if human translations are produced independently of each other (in the same way as variation in less restricted texts, like essays on the same topic). Preliminary results suggest that the degree of lexical distance between two translations is close to a certain norm: the distance score above this norm implies that translations were not produced independently, the distance score below the norm implies that the quality of one of the translations is low. (A human reference translation is usually more distant from MT than from another human translation). In this experiment a corpus of multiple student translations of the same STs will be collected, which will allow establishing norms of variation for different lexical distance measures, and the point where the variation can reach saturation. (E.g., the dynamics of growth for the BLEU or another lexical distance score for a human translation, when more and more human reference translations are added to the reference set – can indicate what is the level of saturation). This experiment will try to establish if this variation can be attributed to communicative excessiveness, and if statistical parameters of communicatively excessive and communicatively relevant information are different. If such parameters are identified, they will become the basis of an evaluation measure for MT that does not require human translations to produce the scores.

MT translation corpus will be matched with the corpus of multiple human translation of the same STs, in order to find the limits when MT goes beyond the norms of legitimate variation. The consequences of the deviations from variation norms will be examined: in what cases this causes a fluency disruption, or a factual error / factual omission. The experiment is aimed at increasing the accuracy of MT evaluation by concentrating on communicatively relevant aspects in the text.

2.4. Measuring the performance of an IE system on template filling tasks for MT output
Precision, Recall and F-scores for the MUMIS IE system for original English text will be compared to the scores for English MT output of different MT systems. The cases of failure in filling IE templates will be examined to establish if they are related to disruption of communicatively relevant information, to any specific kind of MT errors (factual errors and omissions or deviations from conventional style). The cases of success in filling the IE templates will be examined to establish which variation in translation still does not disrupt IE performance, and whether the successful segments are within limits of legitimate variation, found in human translations (the variation can be measured by various lexical distance scores). This experiment will address the question if there is a clear borderline between factual and stylistic errors in MT, and at which point stylistic disruptions distort comprehensiveness of text or result in factual errors.

2.2.3. Integration of the IE annotation and IE evaluation metrics into MT systems

Combining IE-guided source text analysis and IE-based evaluation of the target text. Evaluation results can enforce re-analysis of the source segment or choose the best possible analysis from several candidates.

We will examine how the results of the 1st and 2nd sets of experiments can be combined and integrated into an MT engine: at which point MT system can accept MT-oriented IE annotation for improving the analysis of the ST and in which form the results of the TT evaluation can be used. Since the suggested experiments 1.2 and 2.4 use the same material – football match reports processed by the MUMIS IE system, they can be the basis of this experiment. Different translation candidates will be produced by different MT systems, and the translation candidates for each segment will be evaluated by different IE-oriented evaluation procedures in order to find optimal translations for the texts. We try to make MT-oriented IE annotation of the ST more flexible, allowing different ways of translation, depending on evaluation results for different translation candidates.

II. In the second stage of our project we will develop a prototype IE system, which will produce an extended MT-oriented annotation for the STs, which will be aimed at improving the quality of MT. Extensions to the IE annotation will be based on interpreting IE tasks in the speech act framework (see *section 2.4.*).

2.2.4. A prototype: an MT-oriented IE system

The prototype will produce MT-oriented annotation that extends standard IE annotation in the following way: both illocutionary and perlocutionary effects of the source text will be represented in the annotation. Nowadays standard IE techniques roughly model perlocutionary effects of texts (extralinguistic results of text processing, e.g., the changes in knowledege about some events). On the contrary, MT systems need to use annotation of illocutionary effects of the text (i.e., recognition of communicative intentions of the speaker, which does not necessarily affect extralinguistic knowledge. E.g., if the information is uncertain, the recipient may recognise the speaker's communicative intention, but chose not to change his knowledge about the situation). Illocutionary effects can be annotated by statements in some formal language (like SQL). These SQL-type expressions must be checked, whether or not they contain reliable information (and receive some "confidence scores"). Then the reliable statements can be "executed" within the database in order to model perlocutionary effects of the text (updating recipient's extralinguistic knowledge). But SQL-type statement with both "reliable" and "unreliable" information, as well as negative statements (which are often ignored by classical IE template filling algorithms), can be used to guide MT systems, because all of them represent certain communicatively relevant intentions of the speaker that are intended for translation. So we suggest that mapping from a natural text into the SQL-type statements models "illocutionary" effects of the text, and "executing" these statements in the template database represents the text's "perlocutionary" effects. The resulting annotation should be usable for guiding MT systems to identify communicatively relevant information in STs and to avoid factual errors and omissions.

(The speech act framework also suggests the second direction of extending IE annotation for MT purposes: new types of speech acts can be recognised. Classical IE systems normally process "assertives" that are typically present in news reports, but on some future stages IE systems can be extended to process other types of speech acts, e.g., those recognised in [Searle, 1969]: "directives", "commissives", "expressives", "declarations", as well as indirect speech acts. This extension and its effects on MT will be a possible follow-up of the PhD project).

Parallel multilingual corpora of football match reports (from the FIFA and UEFA sites) and the IE resources from the Sheffield University NLP group (GATE and MUMIS) will be used for the development of the prototype IE system. Texts in the following languages will be used: English, French, Spanish, German, and Russian.

The development of the prototype will involve the following stages:

1. A development part of the corpus will be aligned on the word level.
2. MUMIS templates generated from English texts will be used to develop the aligned corpus of SQL-type statements. The corpus of statements will be produced semi-automatically, and will model illocutionary effects of corresponding fragments in the texts.
3. This corpus will be used for developing an example-based system that maps from natural texts into sets of SQL-type statements.
4. The performance of the system will be evaluated on a testing part the corpus.

This SQL-type annotation will be designed to guide MT systems to correctly identify communicatively relevant information in the source text (similarly to NE annotation). Ways of using this annotation for MT will be specified and expected improvement of MT quality will be estimated.

Future work in this direction (in follow-up projects) will involve proper integration of the IE annotation and IE evaluation measures into MT and comparing the baseline and IE-guided performance of MT systems. In that stage, an access to the source code of MT systems will be necessary.

2.3. Theoretical framework: Communicative excessiveness in translation

Communicative excessiveness (i.e., the property of a text where some linguistic representations do not have illocutionary effects – do not signal extralinguistic communicative intentions of the speaker) often causes factual errors and omissions in translation, becoming the basis of different types of ambiguity. In many cases it is not possible to adequately translate all the information expressed in the source text, so a human translator or an MT system often needs to make a choice, which information is essential for the author's communicative intention and so needs to be translated. It is often difficult to make this choice (even for a human translator) because communicatively "active" information can be expressed in unexpected ways and can even change in different situations of the text's use (e.g., new meanings can be found in the text if the knowledge about the situation changes, or new experience of an individual or a society can lead to discovering new interpretations in some works of classical literature). Human translators meet this difficulty mainly in translating fiction and poetic texts, where the density of communicatively important information is very high. However, communicative excessiveness is present in texts of any genre and is likely to cause factual errors and omissions in MT. The following (frequently quoted) translations of H.Heine's poem into Russian by M.Lermontov and F.Tütchev can illustrate the concept of the communicative excessiveness in a poetic text:

Ein **Fichtenbaum** steht einsam im Norden auf kahler Höh. Ihn schläfert, mit weißer Decke umhüllen ihn Eis und Schnee. Er träumt von einer **Palme** die, fern im Morgenland einsam und schweigend trauert auf brennender Felsenwand.	На севере диком стоит одиноко На голой вершине **сосна**(*'pine-tree'*) И дремлет, качаясь, и снегом сыпучим Одета, как ризой, она. И снится ей все, что в пустыне далекой, В том крае, где солнца восход, Одна и грустна на утесе горючем Прекрасная **пальма**(*'palm'*) растет.	На севере мрачном, на дикой скале **Кедр**(*'cedar'*) одинокий под снегом белеет, И сладко заснул он в инистой мгле, И сон его вьюга лелеет. Про юную **пальму**(*'palm'*) все снится ему, Что в дальних пределах Востока, Под пламенным небом, на знойном холму Стоит и цветет, одинока...
Heinrich Heine	tr. Mikhail Lermontov	tr. Fedor Tütchev
English translation:	A **Pine-Tree** standeth lonely In the North on an upland bare; It standeth whitely shrouded With snow, and sleepeth there. It dreameth of a **Palm Tree** Which far in the East alone, In mournful silence standeth On its ridge of burning stone. (tr. James Thompson)	

Communicatively relevant information in H.Heine's text is expressed by the grammatical gender of the words "*Fichtenbaum*" (*Pine-Tree* – 'masculine') and "*Palme*" (*Palm Tree* – 'feminine'). For a recipient of the German text the poem is a metaphor of love and parting (in Heine's works it belongs to the cycle of poems "Dreams of a far-away sweetheart"). Both names of trees appearing in Heine's poem belong to 'feminine' gender in Russian. M.Lermontov uses exact lexical equivalents for the "*Fichtenbaum*" ('*Pine-Tree*') and "*Palme*" ('*Palm Tree*'), but this translation fails to generate the communicatively relevant opposition of 'masculine' and 'feminine' morphological gender, which leads to an absence of an original illocutionary effect of Lermontov's translation. (A group of students who did not know Heine's original poem and theoretical discussions about these examples were given this translation and asked: "What is this poem about?" None of the students said that it was about love and parting [Latyshev, 1988: 76]). F.Tütchev replaced the "*Pine-Tree*" with the "*Cedar*", which has 'masculine' morphological gender in Russian. This translation strategy preserves the communicatively relevant information of the original in his translation (this communicative effect is reinforced by the use of the adjective "*юную* (*пальму*)" – '*young; youthful*' (*Pine-tree*) – which normally characterises humans). Though this translation departs from the original lexical meaning of "*Fichtenbaum*" in Heine's text.

This example shows that "active", communicatively relevant information, which is necessary for achieving the desired communicative effect, can be expressed by functional elements, such as morphological gender (normally functional elements and relations are "latent", communicatively redundant, which perhaps motivated M.Lermontov's translation). On the contrary, information expressed by lexical means and explicitly stated in the text with lexico-syntactic constructions (attributive, predicative etc.) can be communicatively redundant, and translators can ignore it without harming the illocutionary effect of the text.

In the following translations of limericks from English into Ukrainian and Russian the explicitly stated factual information about the number of cats ("twenty-two") and the number of ladies ("one") was "sacrificed" in order to preserve the stylistic effect of the poems. (In Ukrainian the word for 'kitten' had to be in plural, to rhyme with the previous line, but this word has four syllables, so too little rhythmical space remains for the numeral, which had to have only one syllable. The word for "seven" denotes the biggest number still having that property). In Russian translation a 'tiger' was replaced by a 'bear' (the Russian word for 'bear' rhymes with the word for 'lady' in some morphological cases):

English original	Ukrainian translation	
There was a young man who was bitten,	Юнака покусали за п'яти	'A young man was bitten at his heals
By **twenty-two cats and one kitten**.	**Сім котів і малі кошенята**	By **seven cats and small kittens**
Cried he, "It is clear	– Я умру! – верещить, –	– I will die! – he cries, –
My end is quite near.	О, солодкая мить! –	O, what a sweet moment! –
No matter! I'll die like a Briton!"	Як уміють британці вмирати!	As British can die!'

English original	Russian translation	
There was **a young lady** of Niger	Улыбаясь, **три смелые леди**	Smiling, **three brave ladies**
Who smiled when she rode on a **tiger**.	Разъезжали верхом на **медведе**.	Were riding a **bear**.
They returned from the ride	Вернулись все три	All three returned
With the lady inside	У медведя внутри,	Inside the bear
And the smile on the face of the tiger.	А улыбка - на морде медведя.	And the smile – on the bear's muzzle

The facts of overlooking communicatively relevant meanings (even stylistic or aesthetic) make the translation excessively literal (word-for-word). Informative texts (news reports, instruction manuals, etc.) have different communicative goals than poetic texts: usually these are not aesthetic or stylistic goals, but "assertive" illocutionary effects (i.e., changing recipient's knowledge about the situation), which needs to be preserved in the target text. However, factual errors and omissions for these texts in human translations and MT have the similar reason: the failure to recognise communicatively relevant information, still translating communicatively redundant structures (e.g., default meaning of morphosyntactic constructions and some lexical items, or etymology of proper names, which can make the translation wrong, influent and literal). The following example illustrates a factual error in MT caused by communicative excessiveness (the translation of the IE example from [Hobbs et al., 1996]):

English original	English into Russian MT (ProMT)	
Salvadoran President-elect Alfredo Christiani condemned the **terrorist killing** of Attorney General Roberto Garcia Alvarado	Сальвадорский Избранный президент Алфредо Чристиани осудил **убийство террориста** Министра юстиции и генерального прокурора Роберто Garcia Alvarado.	'Salvadoran elected president Alfredo Christiani condemned the **killing of a terrorist** Minister of Justice and Attorney General Roberto Garcia Alvarado'

In different contexts the ambiguous construction of the type "*the terrorist killing*" can mean something similar to "*the terrorist was killed*". But in this case the chosen translation strategy causes a factual error, which is due to the attempt to translate communicatively redundant information: the default part of speech value of the lexeme "terrorist" within the genitive-case morphosyntactic construction. The intended illocutionary effect could be achieved, e.g., within the instrumental-case construction: *убийство террористами ('killing by the terrorists')* or the attributive construction (which is less fluent): *террористическое убийство ('terroristic killing')*. In the first case the "semantically interpretable" feature of 'number' should be "sacrificed", as well as the default translation of {N+NP} construction, in the second case – the default part of speech value of the lexeme '*terrorist*'. An MT system has to have good reasons to make these "linguistically radical" transformations, it should be guided to make these changes by the need to convey communicatively relevant information (e.g., "terrorists killed the Attorney General"), which has to be identified and properly annotated in the first place.

We will argue that this information is motivated by extralinguistic factors (a model of an intended speech act). In the general case, ambiguities of this type, as well as other problems related to communicative excessiveness, cannot be accurately resolved exclusively within the "core" morphosyntactic, lexical and semantic analysis. MT systems need an adequate pragmatic (e.g., speech act) framework to systematically deal with this type of problems. Note that information needed to guide the MT system in the example above can be explicitly present in IE template. IE annotation can guide linguistically sophisticated MT systems to locate communicatively relevant meanings. To motivate such use of IE technology theoretically, we can interpret it within a formal speech act framework. The next step then can be to extend an IE system in order to produce some MT-oriented annotation, as the speech act framework requires.

Improving the quality of MT with automatic NE recognition is one aspect of this approach: it provides appropriate translation strategies for communicatively relevant information (the type and structure of strings used as NEs), blocking the translation of communicatively redundant information, e.g., etymology of some NEs, etc.(Note, that etymology of NEs can become communicatively relevant under certain circumstances, e.g. Charles Dickens in his novels often used "talking names" for his characters. The etymology of these names is communicatively relevant and is translated into other languages by substituting the name with another name that has the intended etymology in the target language, but still looks like an English name. This shows that the problem of translating NEs is not limited to only morphosyntactic and semantic issues, but is also related to the speech act framework).

The examples above suggest that the information whether some representation belongs to the "active" or "latent" layers is not present in the text itself, but comes from external sources (in a similar way as other pragmatic meanings, e.g., the conversational implicature). The system needs to choose the relevant layer dynamically based on flexible communicatively guided criteria of success/ failure of intended speech acts, not being strictly linked to text-triggered translation equivalents on a specific information layer (as it is the case for the equivalent-based MT architectures). Human translators are more flexible in making the choice of communicatively relevant layers of information (human translation always is "communicatively guided"), and this flexibility needs to be modelled in MT systems.

The conclusion is that the quality of the equivalent-based MT architectures is inherently limited by the level of communicative excessiveness of linguistic representation in text (a similar observation was made for statistical MT, which is limited by redundancy of natural languages [Wilks, 1994: 113].

Improvements in MT quality in this direction can be based on a speech act framework or an approximation of such framework for a specific domain. IE tasks (scenario template filling, NE recognition, etc.) can be the basis of such approximation in the domains of news reports, etc. It is possible to assess communicative relevance and excessiveness of linguistic representations based on the information from filled IE templates, and to choose appropriate translation strategies for these structures, which will enable the avoidance of some serious factual errors in MT.

2.4. Theoretical framework: Speech act model for MT-oriented IE

A complete model of external pragmatic sources that guide interpretation for metaphoric meanings in poetic texts will require an extensive research in future. However a model of communicative relevance and communicative excessiveness in informative texts (that will be usable for guiding appropriate translation transformations in MT systems) can be based on IE techniques interpreted within the speech act framework, proposed by J.Austin [Austin, 1962] and J. Searle [Searle, 1969, 1979].

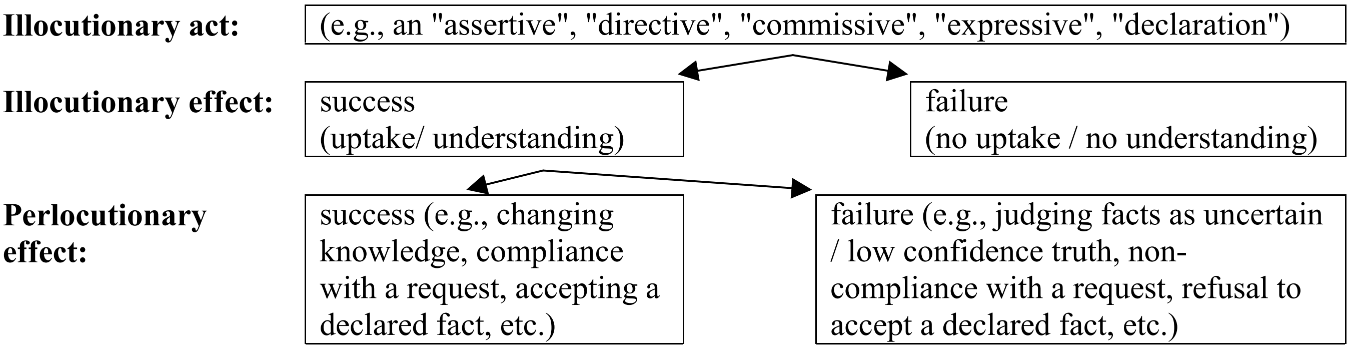
IE is often defined in a very narrow sense as “the automatic identification of selected types of entities, relations, relations, or events in free text” [Grishman, 2003]. However, it is also possible to interpret the scenario template filling task of IE in the speech act framework, as a model of a *perlocutionary effect* of the processed text: e.g., changing system’s knowledge about described events. This interpretation can be productively used in MT and serve as a theoretic ground for improving MT quality with IE techniques, which are able to guide the MT systems in making the distinction between communicatively active and communicatively redundant information and to single out facts and concepts important for translation, avoiding the most serious factual errors.

In the speech act framework, three types of verbal acts are distinguished: *locution* – "the literal use of the utterance with a particular sense" (e.g., the semantic representation of the utterance), *illocution* – which "relates to what the speaker intends to perform" and *perlocution* – which "relates to what is achieved – including uptake by the hearer" [Leech and Weisser, 2003: 139].

If the hearer recognises the speaker's commitment to certain communication task (e.g., an attempt to change hearer's knowledge, to create certain facts: naming something, resigning, etc., to direct one's behaviour: ordering or asking etc.), then the *illocutionary* act is successful (has a successful illocutionary effect). Illocutionary acts are intralinguistic, in the sense that they need to be understood, or taken up, in order to be successful.

Perlocutionary acts are successful if they have intended extralinguistic effect (e.g., if a hearer actually believes in (is certain, or is highly confident of) the statement's truth and changes his/her state of knowledge accordingly, if a hearer accepts the speaker's authority to declare a new fact, complies with the speaker's request, or answers the question).

The following illocutionary acts are distinguished: *assertives* (stating, claiming, reporting, announcing), *directives* (ordering, requesting, demanding, begging), *commissives* (promising, offering, swearing to do something), *expressives* (thanking, apologising, congratulating), *declarations* (naming something, resigning, sentencing, dismissing, excommunicating, christening) [Leech and Weisser, 2003: 140]. The diagram adapted from [Reiss, 1985: 25] summarises the relation between the verbal acts:



Illocutionary act: (e.g., an "assertive", "directive", "commissive", "expressive", "declaration")

Illocutionary effect: success (uptake/ understanding) | failure (no uptake / no understanding)

Perlocutionary effect: success (e.g., changing knowledge, compliance with a request, accepting a declared fact, etc.) | failure (e.g., judging facts as uncertain / low confidence truth, non-compliance with a request, refusal to accept a declared fact, etc.)

The scenario template filling tasks can be interpreted within the speech act framework as a task of recognising perlocutionary effects of "assertive" speech acts, aimed at changing system's knowledge about the described events (news reports, used as the input for IE, rarely have other types of communicative goals).

Firstly, a perlocutionary effect is successful, if a corresponding illocutionary effect (i.e., the recognised speaker's intention to produce the perlocutionary effect, e.g., personal commitment to the truth of some facts, etc.) is also successful. However, some illocutionary effects are not modelled by IE systems, e.g., those which represent uncertain or "not interesting" information, which have no specified template structure, etc. But this information can be still interesting for MT systems: all illocutionary effects are communicatively relevant and need to be adequately translated, but they do not necessarily appear in IE templates. In order to correct this problem MT-oriented IE systems need to distinguish illocutionary and perlocutionary effects, and supply appropriate annotation for the former. This annotation can guide MT system to ensure that the intended illocutionary effects are properly conveyed in the TL. A formal language, like SQL, can serve as annotation for illocutionary effects. This analogy also holds for the case of executing SQL statement: the database records are created or modified by SQL in a similar way as perlocutionary effects are licensed by successful illocutionary effects. Distinguishing formal representations of illocutionary effects (e.g., SQL statements) and IE templates will increase flexibility of IE systems in dealing with MT-specific problems.

Secondly, the speech act framework can extend the flexibility of IE systems to recognise and annotate a wider range of communicative intentions; not only "assertive" speech acts. IE-guided MT systems can use the annotation of illocutionary effects for these additional types of speech acts. This interpretation allows us to develop a more accurate pragmatic model for IE technology that can be used for improving MT quality.

3. The proposed plan and timeline

3.1. Plan of action: Description

1. Literature review – on-going process.
2. Developing experimental resources

   – will include creating and maintaining a multilingual parallel corpus of football match reports and a corpus of multiple students' translations of texts. We will also use DARPA MT evaluation corpus and the DARPA MUC-6 / MUC-7 resources.
3. Experiments (as described above)

   – will include specification of resources, data flow, evaluating and reporting the results.
4. Developing specifications for the IE-guided architecture for MT

   – will include specifying MT-oriented annotation of the ST for IE systems, evaluation procedures for the TT, and ways of using the annotation and the evaluation procedures by the MT system.
5. Developing a prototype MT-oriented IE system

   – we will use the parallel corpus of the match reports and the available development resources of the GATE and MUMIS systems.
6. Analysis of the results
7. Writing up the thesis

3.2. Outline of thesis containing chapter headings and major section headings

Chapter 1. Introduction: IE-guided architecture for MT

- Theoretical foundations for IE guided architecture
- Communicative excessiveness and communicative relevance in translation
- Speech act framework for an MT-oriented IE system

Chapter 2. IE-guided analysis of the source text: IE annotation and use in MT

- NE recognition
- Template element filling
- Scenario template filling
- Co-reference resolution
- Word sense disambiguation
- Summary generation in controlled language

Chapter 3. IE-based evaluation of the target text

- Legitimate variation in translation
- Statistical modelling of MT corpora
- Performance of IE systems on MT output
- IE-based evaluation measures for MT: comparison with other methods

Chapter 4. Specification of the IE-guided architecture for MT

- Annotating illocutionary acts
- Annotating translation strategies
- Using IE annotation in MT
- Evaluation of the output: choice of candidates and re-analysis
- Integrating analysis and evaluation

Chapter 5. Conclusions and future work

Annex. Description of the prototype IE system and MT-oriented annotation scheme.

Proposed timeline

YEAR	03							04												05								
MONTH	06	07	08	09	10	11	12	01	02	03	04	05	06	07	08	09	10	11	12	01	02	03	04	05	06	07	08	09
TASK																												
1. Literature review	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
2. Developing experimental resources	■	■	■	■	■																							
3. Experiments *1)IE-oriented ST annotation*		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■												
2) IE-oriented evaluation of the TT			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■											
3) Integrating IE annotation and evaluation							■	■	■	■	■	■	■	■	■	■	■	■	■									
4. Developing specifications								■	■	■	■	■	■	■	■	■	■	■	■	■								
5. Developing the prototype: *1) Aligning*									■	■	■	■	■															
2) Corpus of aligned SQL statements													■	■	■	■	■											
3) The system for mapping text→SQL														■	■	■	■	■	■	■	■	■	■					
6. Analysis of the results															■	■	■	■	■	■	■	■	■	■				
7. Writing up the thesis																				■	■	■	■	■	■	■	■	■

PhD start date: October 2002. Deadline: October 2005

References

Austin, J.L. 1962. How to do things with words. Oxford, Oxford Unviersity Press.

Babych B., Hartley A. 2003. Improving Machine Translation quality with automatic Named Entity recognition. In: *EACL 2003, 10th Conference of the European Chapter. Proceedings of the 7th International EAMT workshop on MT and other language technology tools. Improving MT through other language technology tools. Resources and tools for building MT.* April 13th 2003, Budapest, Hungary. Pp. 1-8

Babych B., Hartley A., Atwell E., 2003. Statistical modelling of MT output corpora for Information Extraction. In: *Proceedings of the Corpus Linguistics 2003 conference*, edited by Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. Lancaster University (UK), 28 - 31 March 2003. Pp. 62-70.

Berger A. and J. Lafferty. 1999 Information retrieval as statistical translation. In *Proceedings, 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, California, August 1999.

Berger A., R. Caruana, D.Cohn, D. Freitag, and V. Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of SIGIR*, pages 192--199, 2000. http://citeseer.nj.nec.com/303471.html

Carbonell, J.G. and Tomita, M, 1987. Knowledge-based machine translation, the CMU approach. In. Nirenburg (ed.) *Machine Translation: Theoretical and Methodological Issues.* Cambridge: Cambridge University Press, pp. 68-89.

Chesterman, Andrew. 1998. Contrastive functional analysis, Amsterdam, Benjamin, – 224 p.

Gachot D.A., Lange E. and Yang J. 1998. The Systran NLP browser: An application of Machine Translation technology in cross-language information retrieval. In.: Grefenstette G. (ed.). 1988. Cross-Language Information Retrieval. Kluwer Academic Publishers. pp. 105-118.

Grishman, Ralph. 2003 Information Extraction. In: [Mitkov (ed.), 2003: 545-559].

Hobbs J. R., Appelt D., Bear J., Israel D., Kameyama M., Stickel M., Tyson M.1996. FASTUS: a cascaded finite-state transducer for extracting information from natural-language text.

Hutchins, John. 2003. Machine Translation: general overview. In [Mitkov (ed.), 2003: 501-511].

Latyshev L.K. 1988. Perevod: problemy teorii, praktiki i metodiki prepodavanija. Moskow. Prosveshchenije. – 160 pp. (Translation: problems of theory, practice and methodology of teaching. In Russian.)

Leech, Geoffrey and Weisser, Martin. 2003. Pragmatics and Dialogue. In: [Mitkov (ed.), 2003: 136-156]

Macherey, Klaus; Franz Josef Och, Hermann Ney. 2001. Natural Language Understanding Using Statistical Machine Translation". In: *"EUROSPEECH 2001 - 7th European Conference on Speech Communication and Technology", pp. 2205-2208, Aalborg, Denmark, September 2001.* http://citeseer.nj.nec.com/452032.html

Mitkov, Ruslan (ed.). (2003). The Oxford handbook of Computational Linguistics, Oxford University Press, Oxford, New York, – 784 p.

Nirenburg, S. and Carbonell, J.G. 1987. Integrating discourse pragmatics and propositional knowledge for multilingual natural language processing. In: *Computers and Translation* 2: 105-116.

Papineni K, Roukos S, Ward T, Zhu W-J. 2001. BLEU: a method for automatic evaluation of machine translation. IBM research report RC22176 (W0109-022) September 17, 2001

Pazienza, Maria Teresa (ed.) 1997. Information Extraction. A multidisciplinary approach to an emerging information extraction technology. Springer, – 213 pp.

Reiss, Nira. 1985. Speech act taxonomy as a tool for ethnographic description. Amsterdam, Philadelphia, John Benjamins B.V.– 154 p.

Saggion H., H. Cunningham, D. Maynard, K. Bontcheva, O. Hamza, C. Ursu, and Y. Wilks. 2002 Extracting Information for Automatic Indexing of Multimedia Material. In 3rd International Conference on Language Resources and Evaluation (LREC 2002), pages 669--676, Las Palmas, Gran Canaria, Spain, 2002.

Searle, J. 1969. Speech acts. Cambridge, Cambridge University Press

Searle, J. 1979. Expressions and meaning. Cambridge, Cambridge University Press

Somers H., 2003. Machine Translation: latest developments. In: [Mitkov (ed.), 2003: 512-528]

Wilks Y. 1994. Development in MT research in the US. In: *Aslib Proceedings,* vol. 46, no. 4, April 1994. pp. 111-116.

Wilks Y. 1997. Information Extraction as a core language technology. In: [Pazienza (ed.), 1997: 1-9]