Outline of Machine Translation Developments in Europe and America

EAMT and IAMT John Hutchins

PROFILE

W. John Hutchins is the author of articles and books on linguistics, information retrieval, and in particular machine translation - many available from his website (http://www.hutchinsweb.me.uk). He is active in the European Association for Machine Translation (president 1995-2004) and the International Association for Machine Translation (president, 1999-2001).
Principal works: Machine Translation: Past, Present, Future (Chichester: Ellis Horwood, 1986); An Introduction to Machine Translation [with Harold Somers] (London: Academic Press, 1992); Editor of Early years in Machine Translation: Memoirs and Biographies of Pioneers (Amsterdam: John Benjamins, 2000); Editor of MT News International (1991-1997); Compiler of Compendium of Translation Software (2000 to the present) and of the Machine Translation Archive (2004 to the present).

Abstract: This review of developments in machine translation (MT) covers the usage and development of computer aids and systems, production systems for large corporations, Internet aids for individuals, and the wide range of research activity in Europe and North America.

Keywords: Machine translation, Europe, America

Types of systems and tools

Commercial MT systems are widely available in Europe and North America in three basic versions: 'corporate' or 'enterprise' versions for large companies; 'professional' versions for independent professional translators; and 'home' or 'personal' systems for occasional users, e.g. for translating Web pages and emails. The 'home' systems are the most basic type, consisting of little more than the core translation engine. The corporate systems include many additional aids, for pre- and post-editing of texts, for terminology management, for project control, etc. The 'professional' systems provide a selection of those facilities found to be most suitable for translators.

Many large translation services and multinational companies use MT systems for translating large volumes of texts, e.g. in the United States government institutions (DARPA, USAF, etc.) and large corporations (Xerox, Ford, General Motors, etc.). Major users in Europe are companies such as SAP and Siemens, and in particular the European Commission.

Professional translators, translation agencies and smaller companies prefer computer-based translation tools, and in particular translator workstations, often referred to by their most distinctive component as 'translation memory' systems - many developed initially by European companies. The most widely used currently are: SDL, Transit, Déjà Vu, MultiTrans, LogiTerm, Wordfast, and ProMemoria. Each offer similar ranges of facilities and functions: multilingual split-screen word processing; terminology recognition, retrieval and management; creation and use of translation memories (bilingual text corpora of previous translations and their originals); and support for all European and many Asian

languages, both as source and target languages. Finally, and not least, workstations provide access to fully automatic translation if and when required.

Most computer aids for translators are developments from MT research. Examples are authoring systems to help writers to compose texts suitable for MT treatment, and systems for producing 'controlled language' texts (i.e. conforming to MT-friendly syntax and to standard terminology) - the pre-processing of texts input to MT systems is an increasing feature of corporate usage. Other examples are systems for error detection and correction (of both input and output texts), and systems for automatic text prediction (e.g. TransType), which provide suggestions for text completion to aid human translators who frequently translate similar technical documents. These developments are closely linked to efforts to improve access and retrieval of terms and texts from translation memories.

One of the most distinctive features of the European scene are translation companies providing localisation of documentation and products - these companies have acquired considerable experience in the use of translation aids and MT systems. Related to this activity is the development of software for the localisation of websites. With the growth of the Internet, many companies offer information about their products and services, which increasingly needs to be made available in other languages. The information has to be updated regularly, and software such as IBM Websphere has been developed specifically for translating webpages as and when required.

Online services, electronic mail

The provision of translation on-line is now well established. The need is for fast acquisition of foreign-language information; and top quality output is not at all essential. Many MT systems are marketed for the translation of Web pages and of electronic mail, and there is great and increasing usage of MT services (many free), such as the well-known 'Babelfish' available on Yahoo. Others include FreeTranslation, Google Translator, Bing Translator, Tarjim, WorldLingo, and many more online services are being added, both for specific language pairs and for the 'major' languages (English, French, German, Spanish, Arabic, Japanese, Korean, Chinese).

The translation of electronic mail has been offered on most PC based MT software, but it is clear the 'ungrammatical' and colloquial language of email demands more specialised software. To meet this need, the Translution company is marketing customized software for translation of company-internal email communication.

Patents

In contrast to the situation in Japan and other Asian countries, the application of MT to European patents has only recently been discussed. Consequently there are only three systems specifically for translating patents: the PaTrans and SpaTrans systems developed for LingTech A/S to translate English patents into Danish; and the APTrans system designed for generating multilingual patent claims from controlled English language input.

The statistics-based SpaTrans system has

been evaluated in comparison with the rule-based PaTrans system developed in the 1990s from the Eurotra model. In general, the output quality compares well with the older rule-based system; however, the major problem remains the occurrence of new terminology, resulting in many 'unknown' words in the output.

There has been increasing research interest in the issues involved in patent translation. Workshops on the topic have been held at the MT Summits in 2005, 2007, and 2009. The particular features of patent texts (long sentences, novel terminology) and the problems of searching for patent claims have been well aired in these conferences and in other contributions. The specific needs of the European Patent Office and of other patent offices have been the subject of a number of papers covering plans for services and projects; and recently, the European Commission has approved support for the PluTO (patent language translation online) project at Dublin City University.

MT research

Until the mid 1990s, most MT research was still based on the implementation of lexical and grammar rules (with translation via an interlingua or at least 'deep structure' representations) in what is now called rule-based machine translation (RBMT). Currently, the dominant paradigms of MT research are corpus-based. In statistical machine translation (SMT), words and 'phrases' (sequences of two or three words) from a bilingual corpus (of original texts and their translations) are aligned as the basis for a 'translation model' of word-word (and phrase-phrase) frequencies. Translation involves the selection of the most probable words in the target language for each input word and the determination of the most probable sequence of the selected words (on the basis of a monolingual 'language model'). Example-based machine translation (EBMT) involves similar alignment of bilingual data, but here the translation units are larger than individual words or short word sequences; input sentences are matched against phrases or clauses (examples) in the corpus, then equivalent phrases in the target language are extracted, and adapted and combined in acceptable output sentences. Both methods make substantial use of large bilingual corpora, but where SMT is based exclusively on statistical correlations, EBMT applies both statistical techniques and linguistics-based methods similar to those of earlier RBMT approaches.

The advantages of SMT are that systems can be developed rapidly, with relatively scarce data, and for languages previously ignored by RBMT. Initial difficulties with languages having rich morphologies and non-European syntactic structures are being increasingly overcome, and SMT systems are now performing as well as (or better than) RBMT systems. Performance and progress is measured by evaluation metrics, such as BLEU, NIST, METEOR, HTER, etc. These metrics enable researchers to monitor the effectiveness of alternative models and processes, to compare SMT systems and to compare them with RBMT systems. However, it can be argued that metrics developed for SMT are inherently biased against RBMT systems, since the quality of the latter is often ranked higher by human judges.

Although SMT research now dominates MT research, the great majority of commercial systems are RBMT systems. Few SMT systems have reached public operational status. The leader has been Language Weaver offering translation systems for Arabic, Chinese, French, German, Persian, Romanian, Spanish, etc. to and from English. Online services are now predominantly SMT-based, e.g. 'Google Translate', 'Bing Translate' (previously 'Windows Live Translator'), 'Babelfish' (now on the Yahoo site).

Collaboration, open source

Probably the most significant development in MT research in Europe is the establishment of the Euromatrix project (based at Edinburgh University). Its aim is the development of open-source MT technologies applicable to all language pairs within Europe, based on hybrid designs combining statistical and rule-based methods. There will be a particular emphasis on languages of new member states of the European Union, and on systems for translating technical, social and legal documentation.

Euromatrix is not the only example of collaboration. A major feature of recent SMT research is the availability of tools such as parsers and evaluation metrics as open source materials. Examples are Moses, GIZA, Joshua, BLEU, NIST, METEOR, all widely used by researchers and thus facilitating not just rapid development but also comparative evaluations and progress. Following these examples, RBMT and EBMT researchers are now also using open source materials - perhaps best known is the Apertium framework, used for systems for Spanish, Catalan, Portuguese and Basque.

A related aspect is the design of systems with multiple engines (different types of SMT systems, e.g. some with morphological analysis, some using dependency representations, some as 'basic' phrase-based systems). In other cases, RBMT systems have combined with SMT systems. The results have been 'hybrid', 'multi-engine' or 'combination' systems (the terminology fluctuates), in which the final translation is derived from a combination of the outputs of each component system.

Speech

The most innovative area of current research is automatic translation of spoken language. The main centres are ATR in Japan, the Carnegie-Mellon University (USA), the University of Karlsruhe (Germany), all collaborating in a project (C-STAR consortium) to develop speaker-independent real-time telephone translation systems for Japanese, English and German - initially for hotel reservation and conference registration transactions. The government-funded Verbmobil project in Germany intended for business negotiations in German, Japanese, and English, has now ended. Speech translation continues to attract much publicity, but few observers expect dramatic developments in the near future. While we can envisage MT of speech in highly constrained domains (e.g. telephone enquiries, banking transactions, computer input) it seems unlikely that automatic speech translation will extend to open-ended interpersonal communication.

Languages

MT software is available from a large number of European and North American vendors and covering virtually all European language pairs. Here we can mention only the most notable (for a full listing see the Compendium of translation software at http://www.hutchinsweb.me.uk/Compendium.htm). Nearly all cover the major European languages (English, French, German, Italian, Spanish), and many of them also translate from less common Languages (Greek, Polish, Russian, Hungarian, Turkish, etc.) and from and into Arabic, Chinese, Japanese, Korean, etc.

In Europe there is particular need for translation tools for languages of the central and eastern states of the European Union, e.g. Czech, Polish, Hungarian, Latvian, Slovenian, Estonian and Bulgarian. There has also been research on systems for 'minority' languages in Europe, such as Basque, Catalan and Galician in Spain and for immigrant languages such as Hindi, Bengali and Gujarati in the United Kingdom.

The interest of the US government bodies focuses on the use of MT in conjunction with information gathering and analysis. Hence, their support for SMT research in Chinese, Arabic, Farsi, Pashto, etc. There are particular problems involving recognition of names, differences of transliteration, and the integration of MT with information extraction, text mining, intelligence analysis, etc. Research on SMT has undoubtedly been stimulated by this governmental involvement.

MT in the future

Machine translation is demonstrably cost-effective for large scale and/or rapid translation of technical documentation, (highly repetitive) software localization manuals, and many other situations where the costs of MT plus essential human preparation and revision, or the costs of using computerized translation tools (workstations, etc.), are significantly less than those of traditional human translation with no computer aids. This usage is growing and will continue to grow with globalization.

For the translation of texts where the quality of output is not important, machine translation is often the only solution. For example, it will always be uneconomic to produce human translations of scientific and technical documents just for general background information and/or specific data. In these cases, MT will increasingly be the only answer. And there are new applications where human translation has never featured: the production of draft versions for authors writing in a foreign language; the real-time translation of television subtitles; the translation of information from databases; the on-line translation of Web pages; the translation of electronic mail; etc.

The Internet will drive changes in the nature and application of MT. What users of Internet services are seeking is information, in whatever language it may have been written or stored - translation is just a means to that end. Users will want seamless integration of information retrieval, extraction and summarization systems with automatic translation. There is now active research in cross-lingual information retrieval,

multilingual summarization, multilingual text generation from databases, and so forth.

Existing systems have been developed for well-written scientific and technical documents and have assumed that texts will be post-edited by knowledgeable users or translators. Internet usage demands systems specifically for the kind of colloquial (often ill formed and badly spelled) messages found in emails and social networking sites. For such tasks, SMT making use of the voluminous data available on the Internet itself appears to be the only solution.

Information about systems and services mentioned in this review can be found in the Machine Translation Archive (http://www.mt-archive.info)

欧州と米国における機械翻訳開発の概要

ジョン ハッチンス

アブストラクト：本稿の機械翻訳 (MT) 開発の概要は、コンピュータ支援システムの利用と開発、大企業向け製品システム、個人向けインターネット支援、及び欧州と米国における広範囲な研究活動を含んだレビューである。

種々のシステム及びツールの型

ここでは、機械翻訳システム及び関連ツールのタイプを分類し、各々の特徴を挙げている。

先ず、３つのタイプに分けている。

①大企業向け、②プロ翻訳者向け、③一般ユーザ向け

大企業向けは、前編集、後編集、用語管理、プロジェクト制御を含んだもので、プロ翻訳者向けは、大企業向けをカスタマイズしたもの、一般ユーザ向けは、コアの翻訳エンジンに限定したもの。

機械翻訳の利用分野として、欧米では、大量のドキュメントを翻訳するために機械翻訳が利用されている。米国では、政府機関（DARPA、米国空軍、等）と大企業 (ゼロックス、フォード、GM 等)、欧州では、大口ユーザの SAP やシーメンスといった民間企業と、特に欧州共同体がその利用部門である。

翻訳家、翻訳会社、小規模企業では、コンピュータベースの翻訳ツール、特に翻訳ワークステーション（翻訳メモリシステムとも呼ばれる）が好まれている。翻訳メモリは欧州企業により開発され、現在最も使われているのは、SDL, Transit, Déjà Vu, MaltiTrans, LogiTerm, Wordfast, ProMemoria それらは何れも類似の仕様で

ある。共通的な機能は、多言語分離ワープロ、用語管理、翻訳メモリの作成と利用、全欧州言語と多くのアジア言語からなる言語対である。

翻訳者支援機能の多くは、機械翻訳の研究から生まれている。例えば、機械翻訳に向いたオーサリングツールや制限言語、前編集は企業で利用されている。あるいは自動テキスト予測で入力テキストの間違いを訂正してくれる。（TransType では翻訳者に類似のターミノロジーを示してくれる。）

欧州翻訳シーンにおける特徴は、ドキュメントのローカリゼーションで長年の実績があることである。この経験は企業において重要となる多言語版の Web ページ管理を必要なときに更新できるシステムに活用されている。IBM の Websphere はその一つである。

オンラインサービスと電子メール

ここでは、オンライン翻訳について述べている。

オンライン翻訳に関する意識あわせが出来つつある。外国語の情報をすばやく獲得することで、その訳質には余り拘らないものとすることである。多くの機械翻訳システムが Web ページの翻訳や電子メールの翻訳を目標にしている。最近は、インターネットの無料翻訳サービスが増大している。

電子メールの機械翻訳は多くの機械翻訳システムで翻訳のターゲットにされているが、文法的にイリーガルな表現が多くカスタマイズが必要になる。このような市場に Translution 社は企業内の電子メール翻訳に特化したビジネスを始めている。

特許

欧州での特許の機械翻訳は日本やアジア諸国と比較してこれまで議論されていなかった。僅か 3 社の機械翻訳しかなかった。RBMT のものと SMT とが比較されてきたが訳質の評価は同じ程度であった。共通課題は特許における未知語や長文の処理であった。特許翻訳への関心が高まり MT サミットでも 2005 年から議論が始まった。欧州特許庁 EPO をはじめ各国の特許庁でのニーズの高まりで種々のプロポーザルが提出されている。最近、欧州共同体は、特許言語翻訳オンライン PluTO プロジェクトをダブリンシティ大学を中心とするコンソーシアムに 3 年間、2 百万ユーロの計画で、4 月から開始した。

機械翻訳研究

ここでは機械翻訳の研究動向を述べている。

1990 年代半ばまでの中心テーマは、語彙と文法による規則ベース機械翻訳 RBMT であったが、現在の中心パラダイムはコーパスベースである。単語レベルの細かい単位での対応関係を研究する統計ベース機械翻訳 SMT と句や節、文レベルを対象とする用例ベース機械翻訳 EBMT がある。SMT はコーパスが揃えば短期間に開発できるため RBMT では考慮されなかった言語に対しても開発が可能となる。現在、SMT は RBMT と同等な水準に成長している。しかし商用システムでは、圧倒的に RBMT が多い。SMT はほとんど存在しないが、そのリーダ Language Weaver が提供してきた。オンラインサービスは SMT が独占的である：Google, Bing, Babelfish などである。

共同作業とオープンソース

ここでは、機械翻訳開発の方法論として、コラボレーション (共同作業) とオープンソースの重要性を述べている。

欧州の機械翻訳プロジェクト Euromatrix は、SMT と RBMT のハイブリッド型オープンソース開発を行った。これは欧州のような多言語間での対等な翻訳が要求される場合に適切な手法であると述べている。ここから、機械翻訳だけでなくパーザや翻訳品質測定ツールが

提供されている。これらを利用してスペイン中心に、Apertium というフリーなオープンソフト機械翻訳がサービスされている。SMT、RBMT の成果からハイブリッドや多重エンジン、組合せシステム等の設計手法が生まれている。

スピーチ

ここでは音声翻訳について簡単に記載している。

日本（ATR）、米国（CMU）、ドイツ（カールスルーエ大学）を拠点とした話者独立の音声翻訳プロジェクトが終了した。識者は、電話問合せ、銀行業務、コンピュータ入力等の非常に限定された用途では実用可能であるとしても、リアルタイムで話者独立な開放型音声翻訳に対して好意的に見ていないことを述べている。

言語対

ここでは、欧米の機械翻訳システムが対象とする言語対について解説している。

主要欧州言語とその他の欧州言語及びアラビア語、中国語、日本語、韓国語等の相互翻訳である。

EU では特に中欧、東欧の言語間の翻訳ツールのニーズが高い。また、マイノリティ及び移民の言語の研究も継続している。

米国政府の関心は、機械翻訳を利用した情報収集と解析にあり、中国語、アラビア語等の SMT に資金を提供している。

将来の MT

ここでは、機械翻訳の将来性について主としてサービス面から論じている。

大量のドキュメントの翻訳時間の短縮というコスト面でのメリットから機械翻訳は必須である。特に、プロ翻訳家に依存せざるを得ない高品質翻訳分野以外の領域で持続的に増大するであろう。

翻訳品質がそれ程求められない領域として、目を通したり、データを得るために利用する科学技術ドキュメントは機械翻訳しかない。新たな分野として、外国語で書いている作品のドラフト翻訳、テレビのサブタイトルのリアルタイム翻訳、データベース情報の翻訳、Web ページや電子メールのオンライン翻訳等がある。

インターネットが機械翻訳の性格や応用を変化させるであろう。Web ユーザが求めるものはどの言語であれ、情報である。翻訳はその手段である。ユーザは言語間がシームレスにつながることを期待している。自動翻訳を介した情報検索、情報抽出、要約等である。これらは現在活発に研究開発が行われている。

このため、従来のような、後編集も含めた正確な翻訳ではなく、言語的な特徴として会話的（非文を含む）な表現の翻訳が必要で、SMT はインターネット自身から言語データを供給できる唯一の道に思えると述べている。

このレビューで記述した詳細情報は、以下の機械翻訳アーカイブから見つけることが出来る。
http://www.mt-archive.info