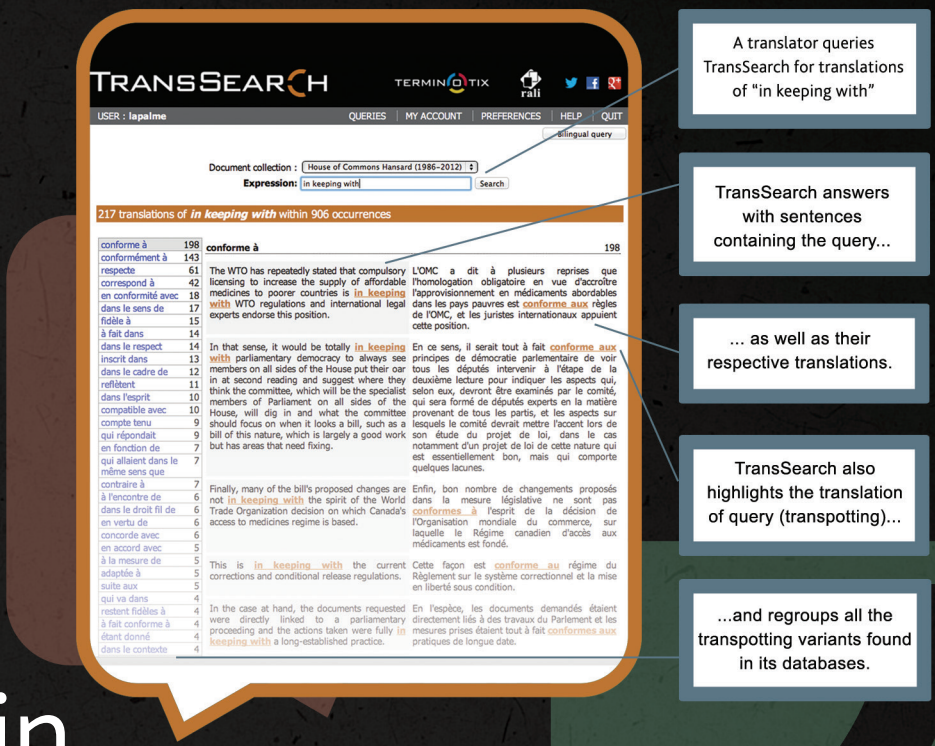
Improving pattern matching in TransSearch



TransSearch is a commercial bilingual concordancer, developed over the last 15 years by **Professor Guy Lapalme** and his team. In this discussion, he explains his interest for translation and how he works to refine the system

To begin, could you provide a description of your position and research interests? How has your background afforded you the expertise to study language translation programmes?

I started to work on Natural Language Processing (NLP) in the 1980s in collaboration with my graduate students. We published the first championship-level programme for Scrabble and were the first to present a computer implementation of a text generator based on the principles of the Meaning-Text Theory, for which we developed a dictionary editor. I was one of the first proponents of logic programming for NLP in Canada.

In 1997, I established the research group Recherche Appliquée en Linguistique Informatique (RALI) for integrating the activities of a computer-aided translation group, which had been operating within the Canadian government since 1984. With three professors, 20 graduate students and postdoctoral fellows plus one researcher, RALI is one of the largest university-based NLP laboratories in Canada.

On top of a strong research activity in statistical machine translation and machine-aided translation, RALI has established close links with industry by means of cooperative research and development grants with Canadian firms such as Terminotix, NLP Technologies, Druide Informatique, Lexum, KeaText, Nuance and Yahoo, and with governmental entities such as Environment Canada. Collaborating with industry on specific problems while ensuring the academic principles are upheld and students successfully complete their degrees is an art that I have managed to develop. This is why I was invited in 2010 to give a keynote speech on this subject at the Canadian AI annual conference.

What other contributions have you made to the field of NLP?

I try to mix innovative research and outreach to the practical world through long-term collaboration with partners from diverse affiliations, both academic and industrial. I have authored more than 300 publications in journals and highly refereed conference proceedings. For my contributions to the field of NLP, I was awarded an honorary doctorate from the University of Neuchâtel and in my academic career, I have supervised 44 Masters, 22 PhD and eight postdoctoral fellows; of these, 18 are now pursuing an academic career as university professors or researchers and three have created NLP-orientated businesses.

How did you develop the tool TransCheck? What is the purpose of this programme?

TransCheck is designed to reduce the cost of quality control by partially automating the job of translation revision. This tool can assist a human reviser by detecting certain types of translation errors and by helping enforce bilingual standards throughout a large translation service.

TransCheck is designed to verify the correspondences between a source text and

PROFESSOR GUY LAPALME

Translating TransSearch

Improved Pattern Matching within TransSearch is a project to enhance a bilingual concordancer with a word alignment functionality. Based at the **University of Montreal**, it aims to find translations of phrases, while providing contexts of occurrence

its draft translation, and to ensure that they maintain certain generally recognised properties of good translation. This is what distinguishes a bi-textual tool like TransCheck from monolingual writing aids like a spelling or a grammar checker. Because they only operate on an uniligual text, the latter are incapable of detecting translation errors, even of the most flagrant kind.

You are currently working on TransSearch. What inspired this project and what do you hope to achieve?

Since 2001, RALI has worked with an industrial partner to commercialise TransSearch, but more importantly, to upgrade the system to take advantage of recent developments in computer science and web-based application design. The research project focused on 'translation spotting', a feature that now allows TransSearch to not only display corresponding sentences, but also highlight the words in each target sentence that are actual translations of the query. This facilitates the work of the translator and broadens the range of applications offered by TransSearch, both on the web and on the user's desktop.

In what ways have you Improved Pattern Matching in TransSearch?

Translation spotting was the focus of research. Two postdoctoral fellows developed, implemented and evaluated numerous algorithms in order to find an appropriate trade-off between precision and speed. Another challenge was merging equivalent translations, ie. translations that are grammatical variations of one another or containing grammatical words that do not change the essence of a translation.

This combination of techniques enabled grouping of identical translations across many examples of the same source text and the computation of occurrence statistics. TransSearch can thus display different translations in decreasing order of frequency, allowing the translator to focus on a few examples already sorted and merged instead of having to pore over several translations.

A GREAT DEAL of time and effort has been invested into improving machine translation, yet experts claim there is still much to explore and improve upon in the field of computer-assisted Natural Language Processing (NLP) tools. One of the latest studies in this field is being conducted by Recherche Appliquée en Linguistique Informatique (RALI), which is one of the largest academic NLP laboratories in Canada. RALI consists of a highly skilled team of computer scientists and linguists with extensive experience in the field. The group's project aims to simplify the workflow of the user by integrating the TransSearch search engine with Improved Pattern Matching, (also known as 'transpotting'), allowing for a corresponding word and its translation to be found within a text.

TransSearch is developed by RALI and led by University of Montreal, Department of Computer Science and Operations Research Professor Guy Lapalme. It is a robust web-based commercial application – a search engine, not dissimilar to Google – aimed at language professionals, designed for finding phrase or word translations among millions of mutually translated pairs of sentences.

At its core, the tool allows translators to submit queries – typically inputted as single words or expressions – to a translation memory, giving access to prepared solutions for a multitude of translational dilemmas. The search engine not only returns sentences containing the query, but also their translation in the other language. Similar translations of the query are grouped, enabling the rapid exploration of various possible translations from which the translator will pick the most appropriate one.

Currently, TransSearch mines translations from more than 500 million words in 15 million English-French sentence pairs of professionally translated texts. The texts are bilingual transcripts of more than 25 years of Canadian House of Commons and Senate Hansards, Canadian Court Rulings and texts from the International Labour Organisation.

Though it does not represent an automatic translator, TransSearch is a translation finder designed to aid translators to make their existing infrastructure more accurate and economical with time. The tool is neither a simple bilingual dictionary nor a thesaurus because it finds translations of entire phrases. Although still lacking the intelligence of a human translator (for example, when addressing complex texts that necessitate high quality analysis and interpretation), it does put the user in complete control for more accurate results.

Numerous factors contributed to the decision to enhance the web-based bilingual concordancer. Research conducted by the laboratory showed that users of the system are, overwhelmingly, professional translators. Furthermore, RALI's studies of user query logs found that the programme is generally used to answer difficult translation problems. Of approximately 7 million queries submitted to the system over a six-year timeframe, 87 per cent contained at least two words. Translators predominantly search for idiomatic expressions, the most recurring of which include: 'in keeping with' (see previous page), 'in light of', or 'look forward to'. Verbs and adjectives containing a preposition were also the object of many queries, including 'consistent with' and 'focus on'.

MODERNISING TRANSSEARCH

The RALI team is striving to update TransSearch, taking advantage of new developments in computing and algorithmic processing of natural language. "As transpotting implied changing the underlying text management system, we also redesigned the architecture to take into account

the new interaction facilities of modern browsers," explains Lapalme. "After consultations with our industrial partner Terminotix, two research associates developed a new user interface that allows a user to query the system, a group manager to deal with subscribers and a system manager to address components." Transpotting, therefore, spurred the team not only to change the foundational text management system but also inspired them to redesign and restructure the system's architecture.

In 2011, the system was handed over to Terminotix, a Montreal-based company specialising in computer-aided translation. Terminotix proceeded to complete some of the essential features, such as localisation, user migration, documentation, and user training. Now in commercial operation, the service has thousands of users, both domestic and international. Bridging the divide between R&D and real world applicability, the successful collaboration between the two parties (RALI and Terminotix) recently led to a multi-year extension to carry the programme forward.

DEVELOPMENT TEAM

RALI's scientists perform both theoretical and practical studies in the area of NLP. Members of the RALI team include Professor Philippe Langlais, who innovates in analogy-based morphology and machine translation, and Professor Jian-Yun Nie, a world-renowned expert in cross-language information retrieval, having recently published a monograph on the subject. They also both have strong teams of graduate students and researchers.

Additionally, RALI has entered a number of internationally renowned competitions and continues to compete in these on a regular basis. Since its establishment more than 15 years ago, RALI launched a series of weekly seminars, inviting researchers and students to present and explain their work.

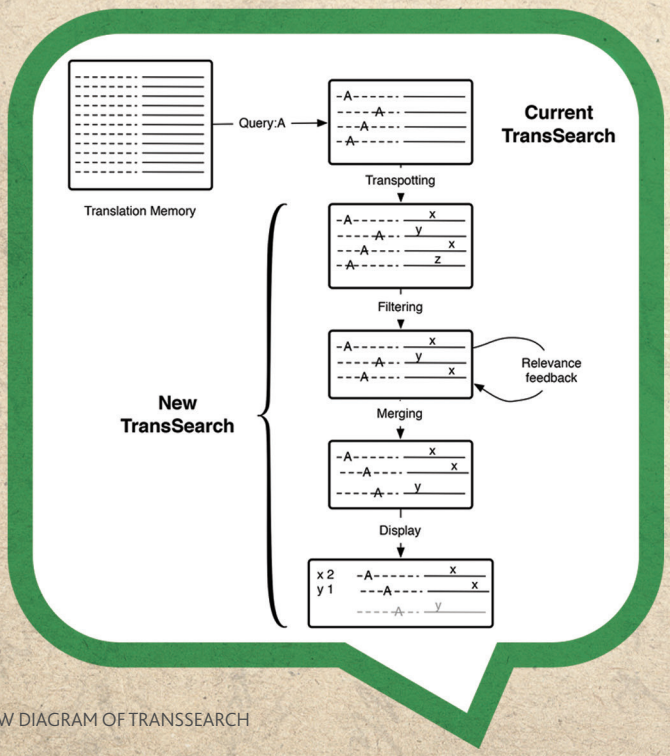
Beyond the computer-aided translation tools featured here, RALI innovated in developing a high quality machine translator of weather warnings issued by Environment Canada.

TRANSCHECK

Another of RALI's recent projects was the development of the prototype for TransCheck, which, as the name suggests, is a tool designed to assist users in detecting translation errors and standardise bilingual translation. The idea behind the venture was to help reduce the cost of quality control by undertaking the job of a human translation reviser or editor. The programme works by aligning the constituent sentences of the source text, as well as the draft translation confirming that the pairs of aligned segments contain no incorrect correspondences and adhere to certain mandatory correspondences.

TransCheck performs four prime verifications:

- Positive terminology: it checks whether terms are translated by an 'authorised translation'. The list of allowed translations can be customised for any application or text
- Negative terminology: using an 'anti-dictionary of discouraged translations', it highlights suspicious translations
- Numerical expressions: a commonly cited mistake – for instance, the transcription of a numerical expression, part number, date, or monetary value – which TransCheck's advanced software can detect. Translating a piece of text that is riddled with numerical expressions can be a very error-prone activity
- Omissions: when the lengths of sentences differ between the source text and translated text, TransCheck will highlight the issue



FLOW DIAGRAM OF TRANSSEARCH

DEFINING ACHIEVEMENTS OF RALI

Several achievements have come to define the laboratory. One such example is the TransType interactive tool, a system that monitors translators as they input text and suggests completions based on those actions, much like predictive text. Each time a key is pressed by a translator, TransType revises its suggestion and provides a new prediction, as Lapalme elaborates: "The novelty lies in the mode of interaction between the user and the machine translation (MT) technology on which the system is based. In previous attempts at interactive MT, the user has had to help the system analyse the source text in order to improve the quality of the resulting MT – using the target text as the medium of interaction results in a tool that is more natural and useful for translators".

During the refinement process, the team encountered a number of technical challenges. For instance, it was essential that the system be designed to restrict suggestions to those that it was reasonably confident with and adapt its behaviour in accordance with the information it had gathered about the human translator. Traditionally, MT only accounts for the source text whereas TransType additionally analyses the

SOME MEMBERS OF RALI

Lapalme is aware that there is still a lot of ground to cover, but is positive that there is a real desire for his translational aids

already inputted target text. Likewise, the tool had to be able to work in real-time and operate quickly enough to match the speed of the user. Another obstacle was ensuring that TransType could make its suggestions and predictions clearly – but not obtrusively – as well as provide an effective visual means for users to accept or decline them.

TransType evolved from the team's collaborative pursuits in enhanced performance, and, through financing from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Quebec Ministry of International Affairs and the European Community, it is hoped that it may one day be operational as a commercial product, delivering a reliable translation system with effective and revelatory suggestions.

OTHER PROJECTS AND FUTURE ENDEAVOURS

The RALI laboratory has been involved in numerous other projects, not least in the field of automatic summarisation – for which it has been a recognised leader since the work, in the late 1990s, of then PhD student Horacio Saggion, now a researcher in Barcelona. The laboratory developed a system, called SumUM that, as Lapalme clarifies: "Produced short summaries of long scientific documents in two steps: first, an indicative summary, which identifies important subjects of a document; then an informative summary, which elaborates on a few subjects selected by the user". Over the past three years, along with PhD student Pierre-Étienne Genest, Lapalme has created an entirely abstractive method of summarisation, despite the majority of investment currently steeped in the extraction of existing sentences from documents.

More recently, Lapalme has dedicated much of his time and interest to opinion and sentiment analysis of texts found on the Internet. With assistance from Marina Sokolova, a then postdoctorate researcher and now professor at the University of Ottawa, Lapalme developed an innovative method of analysing opinion through the study of patterns of unemotional words. Previous studies had generally centred on the distribution of words, specifically adjectives, associated with positive and negative feelings. This new method developed by the two professors, however, is different because it could deal with documents in topics as far-reaching as medicine and insurance, which are traditionally free from emotion.

Up until this point, RALI's work has been aimed at professional translators. However, the group has expressed an interest in exploring other types of users, such as knowledge workers and interpreters who need to compose articles and documents in dual languages. The researchers' experiments and studies have thus far been successful, although more research needs to be dedicated to establishing how the varying needs of users can be integrated into the existing system.

Lapalme is aware that there is still a lot of ground to cover, but is positive that there is a real desire for his translational aids: "In future, we want to continue to further improve our computer tools by combining the expertise gathered during the different projects. For example, now that TransSearch can identify target translations we would like to embed it in a TransType-like interface that would automatically query for possible interesting translations and display them as suggestions for a translator," he concludes.

INTELLIGENCE

TRANSLATION SPOTTING IN TRANSSEARCH

OBJECTIVES

- To focus on the 'identification of translations' (translation spotting), which will display the correspondence between the words of a source sentence and their equivalents in the target sentence
- To facilitate the work of translators and broaden the range of TransSearch applications on the web and the desktop environment
- To allow Canada to maintain its position as a leader in the field of tools for translation

KEY COLLABORATORS

Professors Philippe Langlais; **Jian-Yun Nie**; (Researcher) **Fabrizio Gotti**, Université de Montréal

FUNDING

Natural Sciences and Engineering Research Council of Canada • Fonds de recherche du Québec – Nature et technologies

CONTACT

Professor Guy Lapalme
Principal Investigator

Département d'informatique
et de recherche opérationnelle
Université de Montréal
CP 6128, Succ Centre-Ville
Montréal
Quebec, H3C 3J7
Canada

T +1 514 343 6111 X 47493
E lapalme@iro.umontreal.ca

http://rali.iro.umontreal.ca

GUY LAPALME is Professor of Computer Science at the Université de Montréal, where he has been a faculty member since 1980. He is a world leading expert in the computer processing of human language. He has published on many aspects of the subject including spelling correction, dictionary editing, text generation, automatic summarisation, information extraction, opinion mining and machine translation tools. Lapalme has also made contributions in operations research, compilers and bioinformatics. His career is a successful mix of innovative research and outreach to the practical world through long-term collaboration with partners from diverse provenance, both academic and industrial. Recently, he was awarded an Honorary Doctorate from the Université de Neuchâtel (Switzerland) and Lifetime Achievement Award from the Canadian Artificial Intelligence Association.

TERMINOTIX

Université de Montréal

rali