Mach Translat (2007) 21:139–163
DOI 10.1007/s10590-008-9040-7

Automatic extraction of translations from web-based bilingual materials

Qibo Zhu · Diana Inkpen · Ash Asudeh

Received: 14 September 2007 / Accepted: 7 August 2008 / Published online: 20 September 2008

Abstract This paper describes the framework of the StatCan Daily Translation Extraction System (SDTES), a computer system that maps and compares web-based translation texts of Statistics Canada (StatCan) news releases in the StatCan publication *The Daily*. The goal is to extract translations for translation memory systems, for translation terminology building, for cross-language information retrieval and for corpus-based machine translation systems. Three years of officially published statistical news release texts at http://www.statcan.ca were collected to compose the StatCan *Daily* data bank. The English and French texts in this collection were roughly aligned using the Gale-Church statistical algorithm. After this, boundary markers of text segments and paragraphs were adjusted and the Gale-Church algorithm was run a second time for a more fine-grained text segment alignment. To detect misaligned areas of texts and to prevent mismatched translation pairs from being selected, key textual and structural properties of the mapped texts were automatically identified and used as anchoring features for comparison and misalignment detection. The proposed method has been tested with web-based bilingual materials from five other

Q. Zhu (✉)
Statistics Canada, Ottawa, Canada
e-mail: qzhu5@carleton.ca

Q. Zhu · A. Asudeh
Institute of Cognitive Science, Carleton University, Ottawa, Canada

D. Inkpen
School of Information Technology & Engineering, University of Ottawa, Ottawa, Canada
e-mail: diana@site.uottawa.ca

A. Asudeh
School of Linguistics and Applied Language Studies, Carleton University, Ottawa, Canada
e-mail: ash_asudeh@carleton.ca

Canadian government websites. Results show that the SDTES model is very efficient in extracting translations from published government texts, and very accurate in identifying mismatched translations. With parameters tuned, the text-mapping part can be used to align corpus data collected from official government websites; and the text-comparing component can be applied in prepublication translation quality control and in evaluating the results of statistical machine translation systems.

Keywords Automatic translation extraction · Bitext mapping · Machine translation · Parallel alignment · Translation memory system

1 Introduction

A rapidly growing body of research on natural language processing, machine translation (MT) and translation memory (TM) systems attests to the importance of automatic translation extraction. The applications of extracted translation pairs illustrate significant diversity. If the quality of translations is good and the translation pairs are well-aligned, the translation segments can represent a crucial resource for different natural language processing tasks (Gey et al. 2002; Moore 2002) such as statistical MT (Deng et al. 2006), example-based MT (Hutchins 2005; Simões and Almeida 2006) and cross-language information retrieval (CLIR) (Chen and Gey 2001). Extracted translation pairs can also be a useful reference for translation studies (Neumann and Hansen-Schirra 2005), computer-aided translation (Callison-Burch et al. 2005), and computer-assisted revision of translations (Jutras 2000).

Translation pairs can be extracted from a variety of resources. Some of these resources are unannotated transcriptions of recorded meeting minutes, scanned copies of literary works, and data collections of text-only flat files. Some are resources with rich metainformation, such as annotated bilingual corpora and bilingual hypertexts. Given the growing amount of bilingual documents on the web, it has come to the attention of researchers in computational linguistics that the web can be explored as a bilingual corpus (Resnik 1999; Chen and Nie 2000). As such, the web is becoming an important source for extraction of translation pairs. Currently there is an increasing need for well-aligned bitexts in NLP, MT and translation studies. For example, for corpus-based MT, good quality translation pairs are in great demand as training data or translation templates. For many TM systems, the amount of previously translated source–target pairs and the speed of generating them are not satisfactory. On the one hand, building annotated and well-aligned bilingual corpora to aid MT and human translation is time-consuming and difficult. On the other hand, we have so many officially translated pages on the web that are not explored for the purpose of bilingual corpus building. What we need is to seek ways to mine the widely available web resources to bridge the gap in the scarcity of bilingual data resources. Automatic translation extraction from web-based materials is one of the efforts in this direction. One obvious advantage of automatic translation extraction systems is the potential of enriching currently available electronic translation resources with limited manpower and lexicographical expertise (Tufiş et al. 2004). In this paper, we describe one

such translation extraction system: the StatCan Daily Translation Extraction System (SDTES).

SDTES is a recently built system at Statistics Canada that attempts to amass, process, filter, and format the translated texts from the web-based StatCan *Daily* data collection. There are two major components in the system: a text alignment component, and a misalignment detection component. For the text alignment component, we designed a two-round procedure in adopting the Gale-Church statistical model for the alignment of web contents (Gale and Church 1991). Without this two-round procedure, using the Gale-Church length model to align the *Daily* HTML texts would have either become impossible or would have generated many false translation pairs. The SDTES misalignment detection component integrates different alignment techniques, and takes advantage of some important textual and structural information in HTML texts. It can help distinguish correctly aligned pairs that are true translations from those pairs that are misaligned and unusable. We developed a two-step cognate extraction method to generate cognate lists that could be used to assist the text alignment process for the purpose of detecting misaligned translation pairs. Results show that SDTES can generate rather clean translation pairs—translation pairs that are of good quality and almost error-free as far as alignment is concerned—because the few problematic translation pairs can be automatically identified and eliminated. The precision and recall of the alignment mechanism and of the misalignment detection mechanism are very high. In addition, the text alignment algorithms and methods are straightforward, robust and can be easily applied to most of the government web materials in Canada where government websites have to present web materials in both English and French. The misalignment detection algorithm can help ensure the quality of aligned translation pairs and greatly reduce the repetitive manual tasks of verifying and proofreading the translation pairs before they become officially endorsed for use in MT systems, in TM systems and in other NLP applications.

Previous work in automatic translation extraction of web-based materials is mostly related to mining web contents as a bilingual corpus and aligning the bilingual texts. These are also the two primary challenges facing systems that extract translations from bilingual websites. As we know, the building of bilingual corpora can be traced back to the earliest translated texts, such as texts on the 196 B.C. Rosetta stone (Véronis 2000b). However, the use of the web as a bilingual or multilingual corpus for language studies, for NLP and for CLIR (Chen and Nie 2000; Resnik and Smith 2003; Sigurbjörnsson et al. 2005; Chesñevar et al. 2006; Li and Yang 2006; Wang et al. 2006) is only in its infant stage. The relatively recent exploration of the web as a bilingual or multilingual corpus was made possible by the rapid growth in the number of web pages, and the availability of vast quantities of web-based translation texts involving many language pairs. Till now, the focus of most of the investigations in this field has been on the discovery and pairing of bilingual sites, domains, HTML documents and pages, although new research is emerging in processing and preparing HTML pages for the actual extraction of translation pairs when bilingual web pages are downloaded. One newly-developed approach (Sánchez-Villamil et al. 2006) in this line of research is to classify XHTML tags into categories, assign different substitution costs to these categories of tags, and use the information to aid the splitting and alignment of texts. This method has produced alignment results that are clearly better than the

alignments obtained when they removed all the tags before the texts were aligned. At the same time, extracting translations, whether from unannotated data resources or from metainformation-rich content, inevitably involves methods of aligning bilingual texts. To this end, various influential text alignment approaches (Brown et al. 1991; Gale and Church 1991; Kay and Röscheisen 1993; Wu 1994; Chen 1996)—whether statistical, lexical or hybrid—were developed in the 1990s. However, most of the classical aligners do not include ways to handle the particular features of HTML texts, a text format that is increasingly popular. As a matter of fact, some features of HTML texts can be as important as the length of the sentence in the text alignment process. In this paper, we intend to investigate, through the description of the SDTES system, the link between research in exploring the web as a bilingual corpus and research in the traditional parallel alignment approaches. We are interested in (1) how to mine web content and prepare HTML files for bilingual text alignment; (2) how to adopt one of the most influential alignment approaches, the Gale-Church statistical approach, for the efficient alignment of the web-based bitexts; (3) how to integrate different alignment strategies with the structural features of HTML texts to aid the text alignment process and the automatic misalignment detection operation.

This paper is organized as follows. Section 2 gives a description of the StatCan *Daily* data sets used in SDTES. Section 3 discusses the SDTES methodology. It includes the preprocessing of the data, using the length-based statistical model to align the text segments, cognates extraction, detection of regions of possible alignment errors, and formatting of the aligned texts. In Sect. 4, we highlight the SDTES results and the methods used for the evaluation of the results. Section 5 is the conclusion.

2 Data sets

The currently collected StatCan *Daily* data sets include English and French texts of *Daily* news releases of Statistics Canada from 2004 to 2006. *The Daily* is the flagship publication of Statistics Canada, and arguably the most important document collection of the agency. It is published officially at http://www.statcan.ca every working day, in two official languages. Usually Statistics Canada publishes an average of 5 *Daily* text releases in a day. The file names of the *Daily* HTML pages bear information about the language and date of publication. Release texts can contain expository texts, numerical tables, notes, graphs and note-to-reader text chunks. Some important major releases are long, running for pages, and some regular minor releases are short, containing only a few lines of texts.

Producing *The Daily* is a cooperative effort. There are people providing draft contents, people assembling and processing the texts and tables, people generating charts and graphs, people editing and polishing the texts, people checking the contents and styles, and people disseminating the finished documents. As there are many people working on *The Daily* on a daily basis, there are guidelines for the submission of release articles, templates for the recurring releases, rules for the usage of HTML markup, and procedures for quality-checking. Because there are dedicated professional editors working on the editing and translation of *The Daily* texts, the translations are mostly standard and consistent translations that are in agreement

ENGLISH
 <B><A HREF="td060614.htm"> Wednesday, June 14, 2006 </A> </B> <H2>Monthly Survey of Manufacturing </H2> <table width="450"> <tr> <td>April 2006</td> <td nowrap ="nowrap" align="right"> Previous release </td> </tr> </table> … <H3>Note to readers</H3>
FRENCH
 <B><A HREF="tq060614.htm">Le mercredi 14 juin 2006 </A></B> <H2>Enquête mensuelle sur les industries manufacturières </H2><table width="450" > <tr> <td>Avril 2006</td> <td nowrap="nowrap" align="right"> Communiqué récédent</td></tr> </table> … <H3>Note aux lecteurs</H3>

Fig. 1 HTML source showing that the HTML structures in the two languages are very close to each other

with the official publication guidelines for government departments and agencies in Canada. The HTML documents are formatted according to the common 'look and feel' publishing standards for Canadian government websites. Normally, we can expect to find many correspondences in structural features like the use of cognates and HTML markup (Fig. 1) in the English and French files.

However, because of the writing conventions of different languages, there are sometimes structural differences in expressing the same content. As a result, this can bring about differences in the use of HTML codes. For examples, in French, some characters in texts (such as the letter *o* in the French abbreviation n^o) are in the superscript text format while they are not in that text format in English. In addition, because *The Daily* is also published in a PDF version, some charts, tables and text blocks can be put in floating positions to avoid graphs and tables crossing over PDF pages. This means that the floating elements such as graphs and tables can end up in different text positions in different languages. To keep the contents and particularly the positioning of the contents the same for the two publication versions (PDF and HTML), the floating nature of some charts, tables and text blocks in the PDF version is reflected in the corresponding HTML pages. For example, in Fig. 2, the text block titled "Note to readers" in English is positioned after the second paragraph, but the corresponding French text block under the title "Note aux lecteurs" is right after the first paragraph. The float of this text block disrupts the order of paragraphs, and thus makes the alignment task significantly more difficult.

3 Methodology

The SDTES text alignment approach relies mainly on the length-based statistical translation alignment algorithm of (Gale and Church 1991). To enhance alignment accuracy and to eliminate manual intervention before alignment, the same statistical model is applied twice with different paragraph and sentence boundary definitions.

ENGLISH

The sharp drop in shipments boosted the inventory-to-shipment ratio to 1.31 in April, equal to the ratio's recent high set in February.
Uncertainty in the manufacturing sector, due in part to the strengthened value of the Canadian dollar and soaring input costs, have contributed to some weakness in shipment activity. As a result, the trend of the inventory-to-shipment ratio has been on a gradual, upward movement over the last few months.

Note to readers

Non-durable goods industries include food, beverage and tobacco products, textile mills, textile product mills, clothing, leather and allied

FRENCH

La chute marquée des livraisons a fait grimper le rapport des stocks aux livraisons, celui-ci se situant à 1,31 en avril, soit un niveau égal au sommet atteint en février.

Note aux lecteurs

Le **groupe des biens non durables** comprend les secteurs de l'alimentation, des boissons et des produits du tabac, des usines de textiles, des usines de produits textiles, de l'habillement, du cuir et des

Fig. 2 Floating text blocks disrupting the order of paragraphs as seen from the file set 060614a. The floating text blocks have the text "Note to readers" for English and "Note aux lecteurs" for French

For misalignment detection, we developed a two-step cognate extraction method. First, we used the K-vec algorithm (Fung and Church 1994) for the initial word correspondence mapping. Then we designed the Acceptable Matching Sequence (AMS) search algorithm to generate cognate word lists that could serve as anchoring features

for the detection of misalignment segments. SDTES was implemented in the Linux and Windows-Cygwin platforms with Perl and C as the major programming languages. Procedures and steps also included pre-alignment processing and post-alignment formatting.

3.1 Assembling and pre-alignment processing of the data sets

In collecting the data for SDTES, we directly accessed the archive storage area of *The Daily*. The file name scheme of *The Daily* reflects information about the language of the text and the publication date. For example, in the file name "d080108b.htm" *d* stands for ***d****aily* which means it is an English release. If this is a French release, the first character is *q*, representing the French word ***q****uotidien*. 080108 is the date of publication and *b* means that it is the second release for that day. We took only those files that matched the pattern d0[456] for English and q0[456] for French in the file names. This enabled us to extract *Daily* releases that were published from 2004 to 2006. Since every *Daily* file was officially published on the Statistics Canada website, an alternative way to obtain the files is to simply grab the pages from http://www.statcan.ca/. Once all the data files were assembled, the next thing to do was to preprocess the HTML documents to pave the way for the use of the Gale-Church length-based alignment model. In our system, a text segment can be a sentence, two sentences together, a title, or a text unit in a table cell. For all the HTML files, the common header and footer wrappings were stripped. SDTES also does a conversion of some French coding such as converting the HTML code character entities "é" and "é" to é. The objective is to handle HTML representational variants of French accent characters properly and consistently. One important task in the preprocessing stage was to reorganize some of the document's structures. As we have seen in Sect. 2, *The Daily* release texts can contain floating tables, text chunks and graphs. For reasons of PDF pagination, some charts, graphs, tables and text blocks can chop off different numbers of paragraphs in between (see Fig. 2). If texts were aligned as they were, the Gale and Church model could have generated many spurious and unusable translation correspondences. SDTES automatically reorganizes all the images, charts, graphs and tables, and moves the floating elements to the end of the document so that the perturbation of text segments is avoided and the translation units within these floating elements can be extracted.

3.2 Alignment using the length-based statistical model

For the text alignment component of SDTES, we use the statistical alignment model of (Gale and Church 1991). The basic assumption of the Gale and Church length-based algorithm is that there is a strong likelihood that a long sentence in English will correspond to a long sentence in French; similarly a short sentence in English will correspond to a short sentence in French. Roughly speaking, if the average lengths of sentences in French and English are known, it is possible to set up a distribution of alignment possibilities from the sentence length information.

For a text in English to be matched up with a text in French, there are always a number of different possible alignments. Given parallel HTML texts E (for English) and F (for French), an alignment A is a segmentation of E and F into $L_x \Leftrightarrow L_y$ pairs, where each L_x is an English segment, L_y is a French segment, and L_x and L_y are mutual translations. The goal is to find the maximum likelihood alignment, given a pair of texts:

$$\arg\max_A \Pr(A|E, F)$$

The first approximation is that the probability of any aligned segment pair in an alignment is independent of any other segment pair:

$$\Pr(A|E, F) \approx \prod_{(L_X \Leftrightarrow L_Y) \in A} \Pr(L_X \Leftrightarrow L_Y|E, F)$$

The second approximation is that each $\Pr(L_X \Leftrightarrow L_Y|E, F)$ does not depend on the contents of the specific HTML texts, but depends only on the contents of the segments within the alignment, so that the $\Pr(L_X \Leftrightarrow L_Y|E, F)$ in the above formula becomes $\Pr(L_X \Leftrightarrow L_Y|L_X, L_Y)$. If we think that the only features that can influence the probability of the alignment are the length paradigms, given $s_1 = \text{length}(L_x)$ and $s_2 = \text{length}(L_y)$, then we can have: $\Pr(L_X \Leftrightarrow L_Y|s_1, s_2)$ instead. Next, we convert the maximization of the approximation into a minimum-sum problem, and implement the minimization by dynamic programming:

$$\begin{aligned}\arg\max_A \Pr(A|E, F) &\approx \arg\max_A \prod_{(L_X \Leftrightarrow L_Y) \in A} \Pr(L_X \Leftrightarrow L_Y|L_X, L_Y) \\ &= \arg\min_A \sum_{(L_X \Leftrightarrow L_Y) \in A} -\log \Pr(L_X \Leftrightarrow L_Y|L_X, L_Y)\end{aligned}$$

The Gale-Church algorithm accepts input from two text files at a time: one text file for each of the languages. The preparation of the text files involves (1) breaking the text files into lines of words, one word per line and (2) adding two types of place-holding markers: one for the end of the paragraph (`.EOP`) and one for the end of the sentence (`.EOS`). For the application of the Gale-Church model, the actual sentence-ending or paragraph-ending markers such as `.EOS` and `.EOP` can take arbitrary names, but the number of paragraph-ending markers (`.EOP`) in the English file should be equivalent to the number of paragraph-ending markers in French. If the paragraph numbers are different, the program will not proceed. The variance in the number of sentences (or `.EOS`) will not matter as much because in aligning sentences, we can have different matching types such as 1:2, 2:1, 1:0 and 0:1.

It is not uncommon to find a discrepancy in the number of paragraphs in *The Daily* document pairing process. Usually the gap is not great. Nevertheless, as noted above, even a slight difference of one or two can bring the alignment process to a halt. It would be rather time-consuming to manually identify the exact places where the end-of-paragraph markings are missing, or are to be inserted. To tackle this problem, we

defined the two types of boundary marking symbols in the first round of text alignment as follows:

> *.EOA is a paragraph pseudo-boundary marker that indicates a block of text containing, in most cases, a series of paragraphs separated by a main HTML element. In the first round of alignment, an .EOA marked text block is treated as a 'large' paragraph.*
> *.EOP is a sentence pseudo-boundary marker that actually indicates the end of a paragraph. An .EOP marked paragraph, though usually containing more than one sentence, is considered a 'large' sentence in the first round of text alignment.*

In SDTES, a limited number of HTML tags that can be used to mark the beginning of a text block are categorized as the main HTML elements. They include tags such as `title, table, h1, h2, h3`. In the first round of text alignment, SDTES counts these few main HTML elements to see if the pair of text files has the same number of main HTML features in them. If the numbers are the same, the system splits the texts into blocks separated by these feature HTML tags. After this, the system marks these text blocks using the paragraph pseudo-boundary symbol `.EOA`, and marks the original paragraph ending places using the sentence pseudo-boundary marker `.EOP`. If the numbers of major HTML elements are different, the system treats the whole text document as a single pseudo paragraph and the original paragraphs as pseudo sentences. Therefore, the parameters for the first implementation of the Gale-Church algorithm are:

```
align –D '.EOA' –d '.EOP' filename_en filename_fr
```

Recent research has shown that some classified HTML mark-up codes can help guide the initial splitting of text in bilingual text alignment (Sánchez-Villamil et al. 2006). In the first round of text alignment for SDTES, we focused only on a few selected HTML tags that are almost certain to appear in the texts of both languages. An obvious advantage of doing the macro-level alignment based on these main HTML elements is that by introducing the HTML structural information into the statistical method, we can avoid misalignment across text blocks. Even if there are misalignments, they will be kept within a minimum text region. At the same time, by using only a few key HTML tags, we can avoid some problems that can be caused when different HTML tags are used for the same representational effects in the texts of two languages or when some structural tags such as `<tbody>` and `<ul>` are missing in one part of the translation units.

To do the second round of alignment, SDTES automatically reconstructs the English document and the French document from the aligned paragraphs of the output file. When the first round of text alignment has been completed using the length-based model, the alignment results are put in a file ending in .al with translation text blocks containing paragraphs clearly indicated. However, to realign the texts, the aligned pairs have to be separated into two files: one for the English texts, and the other for the French texts. Thus, the texts in different languages were reassembled and new definitions were given to the two types of boundary markers:

*** Link: 1 - 1 ***
With much of this water consumed by industrial and domestic use within Canada's cities, water supply constitutes an important facet of the interconnectivity of rural and urban areas of Canada.</P>
Comme une grande partie de cet approvisionnement en eau sert à des fins industrielles et domestiques dans les villes canadiennes, il constitue aussi une facette importante de l'interconnectivité fondamentale entre les régions rurales et urbaines au Canada.</P>

*** Link: 2 - 1 ***
<P>The study found that one-third of Canada's population live in six very highly urban watersheds. These six watersheds occupy less than 3% of Canada's land area.</P>
<P>L'étude a révélé que le tiers de la population canadienne habite dans six bassins hydrographiques fortement urbanisés qui couvrent moins de 3 % de la superficie du Canada.</P>

Fig. 3 Examples of aligned text segments from 060105b

`.EOP` *marks the end of each aligned text block in the output file that is generated as a result of the first round of alignment. In most cases it contains one paragraph, but in some cases it can contain two paragraphs.*
`.EOS` *is the text segment ending symbol in the second round of text alignment. It marks the end of a text segment which in most cases includes one sentence.*

By reorganizing the text structures on the basis of the paragraph pairing results in the first around of text alignment, SDTES is able to reset the English and French documents to the original text format prior to the initial alignment. The two input files are processed with the newly assigned boundary symbols `.EOP` and `.EOS`:

```
align –D '.EOP' –d '.EOS' filename_en filename_fr
```

Figure 3 contains some examples that are produced in the second round of text alignment in SDTES. They are results of text alignment at a more fine-grained text segment level. We also call the text segment unit that is aligned with its translation counterpart in each of the translation pairs "a bead of text".

3.3 Cognates extraction based on the K-vec algorithm and Acceptable Matching Sequence search

Since Simard et al. (1992) suggested using cognates in the alignment of bitexts, different definitions and ways of extracting and measuring cognates have been proposed (Melamed 1999; Danielsson and Mühlenbock 2000; Ribeiro et al. 2001; Inkpen et al. 2005). In the parallel alignment context, cognates are not necessarily words with etymological ties. They can be identical or graphically similar occurrences in the two languages (Véronis 2000a).

In the current system, we define cognates as words that share a good portion, either interrupted or uninterrupted, of the character string in a reasonably computable search space. This means that cognates should be "more or less like each other in form" (McArthur 1992) and that for two words to be considered a cognate pair, they should occur in approximately the same text region. Since our goal of finding cognates is to

use them to assist parallel text alignment in the process of misalignment detection, we require that two cognate words appear in a reasonable range of segment numbers. If a source word in French appears in paragraph 3, but the target English word is in paragraph 10, even if they are a cognate pair, they are not recruited in our cognate list. This is because the two words in this case are not within a text region of reasonable range and we do not want two widely separated sentences to be aligned as a translation pair.

The first step in the SDTES cognate extracting algorithm is to produce candidate cognate lists using the K-vec algorithm. The goal is to find cognate candidates within an acceptable text region range and limit the number of words to be considered as cognate pairs. The K-vec method was developed by Fung and Church as a means of generating "a quick-and-dirty estimate of a bilingual lexicon" that "could be used as a starting point for a more detailed alignment algorithm …" (Fung and Church 1994). The assumption is that if two words are translations of each other, they are likely to occur almost an equal number of times in approximately the same region in the parallel texts. In trying to find word-to-word translations, K-vec does not require any prior knowledge of sentence boundaries. The algorithm divides the two texts into K pieces and looks for word correspondences in corresponding pieces. It can be preferable to use a technique that does not rely on sentence boundary information to verify the alignment results of an algorithm that is heavily dependent on sentence boundary information, because at times, we can be uncertain about the correct correspondence of sentence boundaries in two texts. Problems usually surface if cognate pairs occur in text segments with corresponding sentence boundary information, but not in the same K piece text area. In going across sentence boundaries, the K-vec technique can capture these problems, and help detect misalignment, particularly massively misaligned text chunks. It can also confirm the correctness of correspondence of sentence boundaries, when text segments are properly aligned. Although the K value can be chosen and adjusted, care should be taken not to make it too large or too small. If K is very large, the total number of words in each piece would be small and we may have the risk of missing translations. If K is very small, the number of words in each piece would be large and we lose the advantage of dividing the text into pieces for the purpose of locating a word and its translation in corresponding pieces. Fung and Church (1994) suggested that K be equal to the square root of the total number of word tokens in the text.

For the implementation of the K-vec algorithm, we used the `K-vec++` package (Pedersen and Varma 2002).[1] This package was designed for applying the K-vec algorithm to finding word correspondences in parallel texts. It is called the `K-vec++` package because the package extends the K-vec algorithm in a number of ways. Using the Perl programs in the `K-vec++` package, SDTES generates a very rough list that might contain cognates for each pair of documents. Figure 4 contains some sample lines of a resultant list. For each line in the list, the three numbers indicate the corresponding K pieces where the French word and the English word can be found, the K pieces containing the French word and the K pieces in which the English word occur.

[1] http://www.d.umn.edu/~tpederse/parallel.html.

production<>strong<>1 7 3	ces<>experienced<>1 1 1
alors<>rebound<>1 4 1	octobre<>several<>1 10 2
enregistrées<>growth<>1 2 2	industrielle<>industrial<>1 2 2
effaçant<>increase<>1 1 7	diffusion<>of<>1 1 19
production<>as<>1 7 4	rénovation<>by<>1 1 6
à<>sales<>1 11 3	détail<>pharmaceuticals<>1 3 2
ou<>slowed<>1 2 1	hausse<>apartment<>1 7 1
obtenir<>for<>1 1 6	croissance<>were<>1 3 3
chiffres<>stable<>1 1 1	autres<>unchanged<>1 2 3
d'appartements<>of<>1 1 19	cansim<>gross<>1 1 3
travaux<>was<>1 1 6	au<>september<>1 4 4
nombre<>air<>1 1 1	secteur<>automotive<>1 11 2
résidentielle<>residential<>1 1 1	rebond<>sector<>1 1 9
d'octobre<>gross<>1 1 3	cependant<>machinery<>1 1 2
les<>site<>1 15 1	pour<>sector<>2 4 9
connexes<>steel<>1 1 1	comptes<>our<>1 1 2
la<>an<>3 22 5	marchandises<>marked<>1 1 2

Fig. 4 Sample candidate pairs generated by the Perl program `kvec.pl` in the `K-vec++` package for the document pair 061221a

These results are further processed in the current system to extract cognate pairs such as *résidentielle <>residential<>1 1 1* and *industrielle<>industrial<>1 2 2*.

In using the `K-vec++` package, we take into consideration the *lcutoff* threshold. The *lcutoff* parameter is a threshold frequency value for a word pair to be taken as a valid corresponding pair and to be present in the output list of K-vec (Pedersen and Varma 2002). We are more interested in cognate pairs that have close to 1:1 or 2:2 correspondence than pairs that have a frequency ratio of, say 100:200. The high ratio matching pairs, when used as lexical clues in text alignment, can be computationally expensive, and indiscriminative as an anchoring feature. For example, if the candidate cognate pair information is *industrielle<>industrial<>1 20 2*, it means that *industrielle* appears in twenty pieces of the French text, *industrial* appears in two pieces of the English text, and in only one corresponding piece, they co-occur. This would indicate that the chance of these two words to be in the same translation unit is very small, and thus this candidate cognate pair does not have much value in helping detect misalignments in the context. On the other hand, if the *lcutoff* value is low, the French word could have a better chance of pairing with the English word in the corresponding K piece. For example, when the word correspondence information is *industrielle<>industrial<>1 2 2*, it would mean that the chance of *industrielle* and *industrial* being in the same translation unit is high. This is also one of the reasons why, in the candidate cognate list in Fig. 4, we see so many noisy pairs that are not cognates. As soon as the rough candidate cognate list is produced, there is a filtering device in SDTES to keep only those pairs that have the last three numbers in the output line (such as *industrielle <>industrial<>1 2 2*) very close to each other. If the last three numbers for each candidate cognate pair in the list are represented by h, i and j respectively, the system considers only those word pairs with $\max(h, i, j) < 10$ and $\min(h, i, j) > 0$. Further constraints are: if $\max(h, i, j) > 3$, then $\min(h, i, j) \geq 0.5 \max(h, i, j)$. Otherwise, $\min(h, i, j) \leq 3$. In this way, spurious correspondences like *rebond<>sector<>1 1 9*

or *pour<>sector<>2 4 9* can be eliminated before candidate pairs of cognates are processed further.

The second step in the extraction of cognate pairs is to apply our pattern matching algorithm, the Acceptable Matching Sequence (AMS) search. An AMS has two non-overlapping substrings that can be matched in the same order in both of the words in a cognate pair. The algorithm extracts two substrings (θ^a and β^b) from a source word, say a French word (W_1), with a length threshold (T) for the two substrings combined. Then it searches for the string sequence that contains the two substrings in the same order in the target English word (W_2). Skipping some characters is acceptable before, after or between the two substrings; we call them "Don't Care Characters" (DCCs). The initial value a in θ^a is set to 0, and b in β^b to T. If a match does not occur, one substring θ^a is increased in length ($a = a + 1$) while the other substring β^b is decreased ($b = b - 1$). The search continues until a two-substring match is found or $a > T$ or $b < 0$. An AMS can be defined as in the following regular expression:

$$W_1 = (.)^* \theta^a (.)^{\{0,c\}} \beta^b (.)^*$$

Here:

$(.)^*$ is a substring of any character combinations. This substring can also be an empty substring.
$(.)^{\{0,c\}}$ is a substring of 0 to c characters. $0 < c < 4$.
θ^a is the first substring to be matched in W_2. The length of this substring is a.
β^b is the second substring to be matched in W_2. The length of this substring is b.

We define x and y as the lower bound and upper bound for the lengths of W_1 and W_2, and z is the length difference threshold. $x \leq length(W_1) \leq y$ or $x \leq length(W_2) \leq y$. $|length(W_1) - length(W_2)| \leq z$; if $y > 10$ then $0 \leq z \leq 4$, otherwise $0 \leq z \leq 3$.

We define T as the combined length of θ^a and β^b. $0 \leq a \leq T, 0 \leq b \leq T$, $a + b = T$. We set the T parameter with reference to the upper bound y for the lengths of W_1 and W_2. We discard word pairs with $y < 4$. We set $T = 8$ if $y > 10$. For all the rest, we use the simple linear regression model $T = 0.5y + 1.8$ to compute the threshold value. The linear regression model is derived from the regression analysis of a hand-picked collection of cognate pairs from the *Daily* release texts.

The AMS search model can be demonstrated as in Fig. 5. For example, if the length of the longer string of two words (W_1 and W_2) is 9 and the length difference between the two words is 3, we set $c = 2$ and $T = 6$. Figure 6 represents one of the 'worst-case' AMS search problems in SDTES. Here, we want to find cognates from two candidate words of which the longer one has a length of more than 10 characters ($y = 13$). According to the criterion for the minimum combined matching length requirement, T is 8 when $y > 10$. The two substrings can be separated by 0 to 3 characters ($c = 3$). The length difference between the two words is less than or equal to 4($|length(W_1) - length(W_2)| \leq z; z = 4$). In the solution to the AMS search problem, we are interested in finding if the properties of the two substructures are shared in both of the words W_1 and W_2. If the two substrings θ^a and β^b are exactly the same in the two words, and they occur in W_1 and W_2 in the same order, we say W_1 and W_2 are cognates. Figure 6 shows the AMS search process.

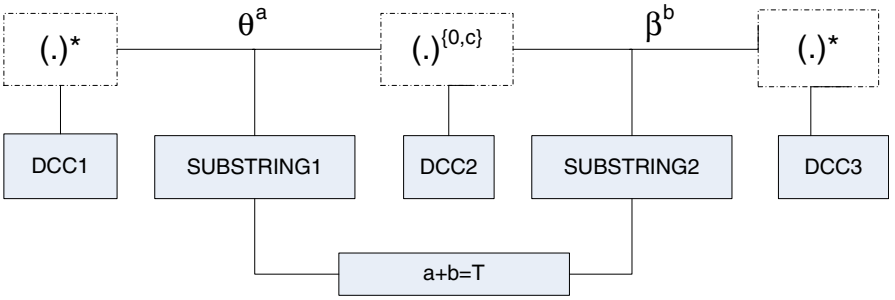

Fig. 5 AMS search model for $x \leq length(W_1) \leq y$. W_1 becomes an AMS only when the two substrings (θ^a and β^b) are both matched in the correct order in W_2. DCC1, DCC2, DCC3 are "Don't Care Characters". They do not need to be matched in W_1 and W_2

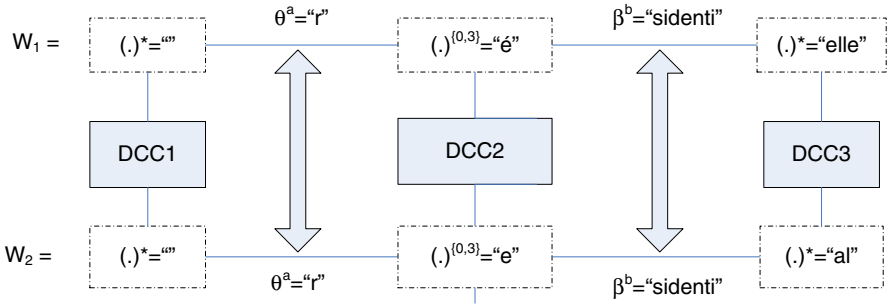

Fig. 6 AMS search process for the candidate cognate pair W_1 = "résidentielle", W_2 = "residential". Here θ^a = "r" and β^b = "sidenti"; $a = 1$ and $b = 7$; c = 3; $T = 8$; $x = 11$; $y = 13$; $z = 4$. W_1 and W_2 are cognates because the two substrings match in the correct order in W_1 and W_2

We wrote Perl code to implement the AMS search algorithm, taking full advantage of Perl's outstanding features of efficient regular expression matching. By applying the AMS search model to the post-processed rough candidate list that was generated by the K-vec algorithm, SDTES was able to produce a cognate matching list (see Fig. 7) for each of the HTML document pairs.

AMS has the straightforwardness of the naïve matching algorithm of Simard et al. (1992). Since in AMS, string matching is conducted at the level of substrings, a substring is treated as if it were a single character unit in the search process. Our goal is to find only the acceptable matching sequence, not necessarily the longest common sequence. Once a two-substring match is found, the search stops. There is no need to do calculations of insertion or deletion to obtain the minimum edit distance. This can reduce the complexity of computational operations. When compared with the string matching algorithm of Simard et al. (1992), AMS adds accuracy (two substrings should match instead of one, avoiding matching problems caused by common prefixes), and flexibility (not necessarily the first four letters should match). This improvement can increase the number of correct cognate pairs identified. At the same time, AMS inherits the strength of the no-crossing-links constraint in the Longest Common Subsequence Ratio (LCSR) algorithm of Melamed (1999). However, in specifying that no more

novembre<>november<>2 2 2	industries<>industries<>2 3 2
cansim<>cansim<>1 1 1	durables<>durable<>1 1 1
centres<>centres<>1 1 1	publications<>publications<>1 1 1
définitions<>definitions<>1 1 1	pharmaceutiques<>pharmaceuticals<>1 2 2
observée<>observed<>2 2 2	résidentiels<>non-residential<>1 1 2
visiteurs<>visitors<>1 1 1	industrie<>industry<>3 3 6
secteurs<>sectors<>4 5 6	totale<>total<>1 1 1
situation<>situation<>2 2 2	manufacturière<>manufacturers<>1 2 2
canada<>canada<>1 2 2	marchandises<>merchandise<>1 1 1
résidentielle<>non-residential<>1 1 2	résidentielle<>residential<>1 1 1
consécutif<>consecutive<>1 1 2	module<>module<>1 1 1
site<>site<>1 2 1	activités<>activities<>1 1 1
produits<>products<>1 3 3	persistante<>persistent<>1 1 1
source<>sources<>1 1 1	qualité<>quality<>1 1 1
l'exploration<>exploration<>1 1 1	machines<>machinery<>1 2 2
base<>base<>1 1 1	tourisme<>tourism<>1 1 1
groupes<>groups<>1 1 1	excluant<>excluding<>1 1 1
mines<>mines<>2 3 2	labrador<>labrador<>1 1 1
	industrielle<>industrial<>1 2 2

Fig. 7 Output cognate list of the AMS algorithm for the document pair 061221a. The input list is the candidate cognate list generated by the K-vec method

than two matching substrings are allowed, AMS overcomes the inherent weakness of LCSR in positing non-intuitive links because of lack of context sensitivity as noted in (Kondrak and Dorr 2004). This can help reduce the number of false positives for SDTES such as v**oi**t**ures/**s**ources,** v**entes/**m**etres**, p**ar**u**t**i**on/**s**tartin**g, m**ensuels/resul**ts, and **courtiers/co**mp**uters**. Also, by focusing on only those pairs that can help with the text alignment, the algorithm is able to filter out many candidate pairs that are either not genuine translation correspondences, or true cognates that could not help the text alignment prior to the actual substring matching process. Therefore, the AMS search model is easy to use and efficient in achieving our purpose.

3.4 Detection of regions of possible alignment errors

As stated in the description of the data sets, translations in the published *Daily* releases are mostly clean and consistent translations in agreement with the Canadian guidelines of publications for government websites. Translation errors such as deletion and insertion are rare. Nevertheless, it is possible that the correct translation pairs are misaligned. In SDTES, we developed an algorithm to detect the regions of possible alignment errors. Figure 8 is the alignment detection operation diagram. It shows the major steps involved in arriving at one of the two outcomes: *pass* and *problem*. The detecting process starts from two prior filtering mechanisms. One of them is the length ratio criteria. If a text segment in one language is more than 3 times longer than the corresponding text segment in the other language, the pair is marked as a problem pair. The second criterion is matching type. Because the extracted translations are independent translation pairs that will be used for MT systems and CLIR systems, matching types such as 1:0 and 0:1 have to be discarded. When these two criteria have

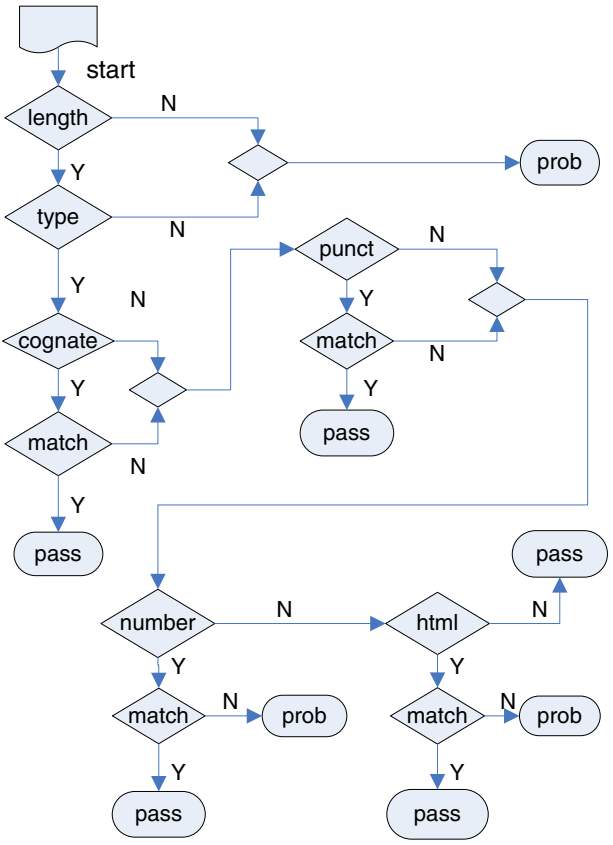

Fig. 8 Text alignment problem detection framework

been checked, SDTES compares the structural and lexical clues of the HTML text segments for further detection. These clues include selected cognates, punctuations, numbers and HTML tags.

For the cognates, we use the list of words that are automatically extracted by the algorithms we described above. When SDTES accepts an aligned pair as input in the misalignment detection operation, it scans through the input text segment of one language to see if there is a word that can be found in the cognate word list. If there is, then the system searches in the corresponding text segment of the other language to see if it contains a word that matches the corresponding cognate word from the cognate word list. If both words in the cognate word list can be found in the aligned pair, the two text segments are a good translation unit. Otherwise, the candidate pair of text segments is passed to the next step of detection. For punctuation, SDTES mainly maps the correspondences using comma, plus sign and minus sign, parentheses, colon and semi-colon, etc. If the system identifies the equivalence in the use of punctuations (frequency of occurrence $n > 0$), it marks the pair as *pass*. In the statistical news release texts, we assume that numbers are translated literally. If we have numbers

```
18 (060628c, 1 - 1) --- pass numbers: full match (e= 20 20 f= 20 20)
19 (060628c, 1 - 1) --- pass cognates: match (e=nationally f=nationale)
20 (060628c, 1 - 1) --- problem numbers: not_match (e= 0 2004 3 5 8 f= 5 8)
21 (060628c, 1 - 1) --- pass numbers: full match (e= 2005 f= 2005)
22 (060628c, 1 - 1) --- pass numbers: full match (e= 20 20 f= 20 20)
23 (060628c, 1 - 1) --- pass numbers: no_numbers,html_tag: no_html_tags,
no_other_clues: match (e= f= )
24 (060628c, 1 - 1) --- pass cognates: match (e=industries f=l'industrie)
25 (060628c, 1 - 1) --- pass numbers: full match (e= 20 20 f= 20 20)
26 (060628c, 1 - 1) --- pass cognates: match (e=all-important f=importante)
27 (060628c, 1 - 1) --- pass numbers: full match (e= 20 20 f= 20 20)
28 (060628c, 1 - 1) --- pass cognates: match (e=alberta f=l'alberta)
```

Fig. 9 Sample of misalignment detection. 'Pass' means the alignment is identified as correct, and 'problem' means the pair is a misalignment

in the English text, normally we should have numbers in the corresponding French text segment. If most of the numbers in the candidate pair do not match, the decision is that there might be a problem in alignment or in translation. When the system does not find any numbers in the aligned text segments, it continues the detection by looking for markups of HTML coding. The distinguishing HTML markups that SDTES uses include <h2> <h3> <i> <b> <a …><table> and others. We did some HTML style unification formatting so that some parts of the HTML codes are highlighted, while some are ignored. For example, the code <a …> becomes <a alink> after the unification formatting. Once the system gathers the key HTML structural features in the aligned texts, a comparison is done to see if the HTML tags are the same. If they are, the segments are a good aligned pair of translations; if they are not, the system marks it as a problematic alignment region. Finally, we have the default-tolerance principle: if there are no structural and textual clues present, and if the two prior filtering criteria (length and matching type) are checked, we mark the segment as *pass*. A sample of the misalignment detection results list is provided in Fig. 9.

3.5 Post-alignment filtering and formatting

We added a filter on the list of stored translations to eliminate aligned pairs such as:

1. Pairs that contain pure metainformation coding or codes that are derived from the HTML coding unification process.
   *** Link: 1 - 1 ***
   <hr size=1> </zcorpus text=10> <zcorpus table=10>
   <hpara=e35:f35>
   <hr size=1> </zcorpus text=10> <zcorpus table=10>
   <hpara=e35:f35>
   (040107a)
2. Alignment segments that include only the numerical information:
   *** Link: 1 - 1 ***

```
2001/02 <hpara=e35:f35>
2001-2002 <hpara=e35:f35>
(040107a)
```

3. Only one pair is kept for duplicate or similar patterns and constructions that are frequently seen in the collection of texts. Sometimes this involves unifying or discarding some information such as numbers and tags in the text.

```
*** Link: 1 - 1 ***
Retail trade</TITLE>
Commerce de détail</TITLE>
(060821a)
*** Link: 1 - 1 ***
<p><B>Definitions, data sources and methods:
survey numbers, including related surveys,
<a alink> 2406 </A> and <a alink>2408</A>.
</B></P>
<p><B>Définitions, source de données et méthodes:
numéros d'enquête, y compris ceux des enquêtes
connexes, <a alink>2406</A> et <a alink> 2408
</A>. </B></P>
(060821a)
```

These aligned segments were further cleaned and organized in the XML format for easy exportability into the TM system, IR database systems, and other application systems. Meta-information items about each of the aligned pairs were recorded, including the string length information (before the HTML codes are stripped), the source of the matched strings, the matching patterns (1:1, 1:2, 2:1, 2:2), and misalignment detection result (*pas* for *pass* and *pro* for *problem*). The SDTES output format (see Fig. 10 for a modified version) is intended to serve as an intermediary form from which different user-friendly formats for different applications can be derived. The interim format is good for use with UNIX system tools and Perl short programs for quick finding of translations, as well as for various sorting, analyzing and statistical tasks. It can be easily converted into different feeding formats in different systems such as *Daily* translation recycling templates system (text block-based format), TM systems (format with texts of different languages assembled in different files), bilingual IR (format required by the search engine), and bilingual text navigation interface (format required by the field specifications of the SQL database). Figure 10 shows the basic XML output format of some final aligned segments with the English part beginning with 'Wholesale trade activity'. For the sake of presentation clarity, line breaks are added to different levels of XML elements. There is also a program in SDTES to convert the basic XML output format to the standard TMX format with a view to expanding the applicability of SDTES to other different situations.

4 Evaluation and results

To evaluate the performance of the aligner, we need a reference collection of manually aligned text segments. The aligned pairs are supposed to be correct in this reference

```
<bead>
<en>Wholesale trade activity declined 1.4% in July, dragged down by reduced sales of
computers and other electronic equipment, lumber and millwork, personal and household
products and oil products.</en>
<fr>Le commerce de gros a reculé de 1,4 % en juillet, freiné par la contraction de la
demande d'ordinateurs et d'autre matériel électronique, de bois d'oeuvre et de menuiseries,
de produits personnels et ménagers ainsi que de produits pétroliers.</fr>
<pa>1:1</pa>
<id>050930a:36</id>
<re>Pas</re>
<le>194=243</le>
</bead>
<bead>
<en>Wholesale trade activity increased 0.9% in June, helped by the demand for
computers and other electronic equipment. </en>
<fr>Le commerce de gros a augmenté de 0,9 % en juin, profitant de la demande
d'ordinateurs et d'autres équipements électroniques.</fr>
<pa>1:1</pa>
<id> 050831a:64 </id>
<re>pas</re>
<le>116=126</le>
</bead>
<bead>
<en>Wholesale trade activity increased 1.0% in October. </en>
<fr>Le commerce de gros a affiché une croissance de 1,0 % en octobre. </fr>
<pa>1:1</pa>
<id>051223a:28</id>
<re>pas</re>
<le>52=66</le>
</bead>
```

Fig. 10 Modified SDTES format for aligned segments for English texts beginning with 'Wholesale trade activity'. In the original SDTES format, only the tag </bead> takes line breaks

collection. Then we can measure how the machine-aligned pairs differ from those in the reference collection.

Let M be the set of segments in the manually aligned reference collection, and A be the set of machine-aligned segments before the misalignment detection device is applied. Precision (P) and recall (R) can be defined as follows:

$$P = \frac{|A \cap M|}{|A|} \quad R = \frac{|A \cap M|}{|M|}$$

For evaluation, we randomly chose one aligned file for every 2 months. The same file is manually aligned for the building of the reference collection. Then we checked the machine-aligned pairs to arrive at the number of correctly aligned segments. For the 18 aligned files used for evaluation, the average alignment precision for aligned translation pairs before filtering is 0.96; and the recall is also 0.96 (see Table 1). This compares favorably with the precision and recall values of most of the other alignment classifiers.

Table 1 Evaluation parameters for aligned text segments before filtering

File	Machine proposed matching pairs before filtering (A)	Manually identified matching pairs (M)	Intersection of A and M	Precision (P)	Recall (R)
040213b	214	214	212	0.99	0.99
040415a	109	109	103	0.94	0.94
040611b	92	94	83	0.90	0.88
040715e	17	16	15	0.88	0.94
041015c	11	11	11	1.00	1.00
041223a	100	100	100	1.00	1.00
050107a	119	119	119	1.00	1.00
050414a	105	105	105	1.00	1.00
050629b	72	72	72	1.00	1.00
050721a	121	120	119	0.98	0.99
050922a	116	116	116	1.00	1.00
051109a	185	190	185	1.00	0.97
060111a	86	87	84	0.98	0.97
060308b	48	48	48	1.00	1.00
060608d	31	34	29	0.94	0.85
060822b	57	58	48	0.84	0.83
061005b	117	117	117	1.00	1.00
061213e	17	16	15	0.88	0.94
Average				0.96	0.96

The main function of the misalignment detection algorithm is to filter out the misaligned translation pairs. The task is to traverse every pair of aligned segments to see if they are indeed a correctly aligned pair. It can happen that the aligned pair is a perfect translation of each other, but the detection algorithm wrongly labels the pair as a misaligned one, or a misaligned pair can be detected as a pair of perfect translations and the algorithm fails to capture the misalignment. In evaluating the adequacy of this filtering component, we used the same files that were randomly chosen for the evaluation of the aligning algorithm. The system derives the machine-proposed alignments (A') by excluding those translation pairs that the filtering device identifies as misaligned segments. M is the reference set, i.e. the number of aligned translation units that the human evaluator thinks the system should have reported. The recall (R) represents the proportion of algorithm-proposed translation units (A') that are right with respect to the reference (M), and the precision (P) is the proportion of correctly proposed alignment segments with respect to the total of those proposed (A').

$$P = \frac{|A' \cap M|}{|A'|} \quad R = \frac{|A' \cap M|}{|M|}$$

From the results shown in Table 2, we can see that the system is accurate in identifying correctly aligned translation pairs (P = .99 and R = .95). A comparison of the precision and recall values for the data sets before and after the filtering shows the effect of the filtering performed by the misalignment detection algorithm. Precision improved from .96 to .99 with a negligible loss of recall from .96 to .95. In the context of automatic extraction of translations, where misaligned or doubtful pairs of translations should

Table 2 Evaluation parameters for aligned text segments after filtering

File	Machine proposed matching pairs after filtering (A′)	Manually identified matching pairs (M)	Intersection of A′ and M	Precision (P)	Recall (R)
040213b	212	214	211	1.00	0.99
040415a	105	109	102	0.97	0.94
040611b	78	94	77	0.99	0.82
040715e	15	16	15	1.00	0.94
041015c	11	11	11	1.00	1.00
041223a	100	100	100	1.00	1.00
050107a	118	119	118	1.00	0.99
050414a	105	105	105	1.00	1.00
050629b	72	72	72	1.00	1.00
050721a	121	120	119	0.98	0.99
050922a	116	116	116	1.00	1.00
051109a	183	190	180	0.98	0.95
060111a	84	87	84	1.00	0.97
060308b	47	48	47	1.00	0.98
060608d	27	34	27	1.00	0.79
060822b	49	58	46	0.94	0.79
061005b	116	117	116	1.00	0.99
061213e	15	16	15	1.00	0.94
Average				0.99	0.95

Table 3 Alignment types for the StatCan data collection

Alignment pattern	1:1	1:2	2:1	0:1	2:2	1:0
Frequency	92,495	2997	1307	9	112	22
Percentage	0.9541	0.0309	0.0135	0.0001	0.0012	0.0002

Here aligned text segments may include lines of pure HTML markup or other metainformation codes

be omitted, this trade-off for the purpose of maximizing precision can be a preferred option.

All in all, SDTES assembled a total of 3,874 documents for each language (English and French) to compose the StatCan bilingual data collection. The collected web documents were published over a period of 736 days, with an average of around 5 release texts a day (the actual number is 5.264). This bilingual *Daily* data collection has 1,611,230 running words for English and 2,050,948 running words for French. The word translation ratio of French to English is 1.3:1, which means that for every ten English words, we use 13 French words in the StatCan *Daily* texts. Using the Gale and Church length-based model, all the text segments were aligned with one of the 6 alignment patterns: 0:1, 1:0, 1:1, 1:2, 2:1, 2:2. The most common pattern is 1:1 ($p > .95$) indicating that more than 95% of the text segments are sentence to sentence translations. There are examples of paraphrasing (patterns 2:1, 1:2 and 2:2), but deletions and insertions (patterns 1:0 and 0:1) in translation are rare (see Table 3).

SDTES parsed all the aligned text segments and used key features such as numbers, HTML markups and cognates to judge if there were possible alignment errors in each of the aligned pairs. 96,283 translation pairs passed the alignment check, while a total

Table 4 Examples of false friends in SDTES cognate matching

French	English	French	English
consiste	considers	sport	report
cours	court	estivaux	festival
abordable	affordable	fiscale	scale
aller	smaller	finlande	mainland
exercer	exerted	grands	brands
variable	available	mains	gains
lever	every		

of just 659 problematic aligned segments were detected. The number of misaligned pairs that escaped detection was minimal. Although there were examples of false misalignment, 99% of the identified correct alignment pairs are truly reliable translation pairs. After filtering and formatting, 70,555 aligned segment pairs were generated. The SDTES model for translation extraction and misalignment detection has been tested with bilingual materials from five other Canadian government websites. More than 200,000 translation units were produced with a level of accuracy and efficiency that is equal to or very close to the level for the SDTES output. These translation units are ready to be used in many applications and systems such as TM systems, IR templates, NLP databases, and MT systems.

While our methodology has achieved good outputs for the officially published government bilingual text data, there are also limitations. When we examine the aligned pairs before the filtering device is applied, we still find some (albeit a very small number of) examples of chains of misaligned sentences caused by the swapping of positions in the translation texts. For example, if in a translation document pair, the 35th text segment in English (L_{X35}) is the translation of the 40th text segment in French (L_{Y40}), although the translation is right, it can amass misaligned segment pairs because the translation spans beyond more than 3 text segments, and the Gale and Church algorithm cannot handle it.

Although the StatCan *Daily* release texts in the bilingual text collection are suitable for use by the Gale-Church algorithm, some short sentences on only one side of the alignment pairs which carry no discriminative feature information can be a source of misalignment. Judging by the length-based alignment criterion, appending the short sentence to the previous sentence or combining the short sentence with the following sentence would not make much difference. This increases difficulty in alignment.

When examining the SDTES-generated cognate lists word by word, we found some misclassified pairs (see Table 4). The false friends are mostly false positives caused by accidental similarity in the orthographic form. They are likely to be on the lists of other cognate classifiers that depend on measures of orthographic similarity.

Although these false friends can potentially hurt the misalignment detection algorithm in SDTES, their impact on the actual identification of misaligned pairs is not so significant. We extracted all the lines which match "…pass cognates:" in the misalignment detection result file, and discovered that not many false friends are actually used for the cognate matching criterion in the misalignment detection process. Generally speaking, if two cognate words are not true friends, and if the text segments where they

occur are not true translations of each other, chances are good that the misalignment would have been detected by the length criterion and the matching type criterion prior to the application of the cognates matching process (see Fig. 8). In some cases, there are other cognate pairs that are checked before the false friend pair. Once a cognate match is found in one of these candidate cognate pairs, the aligned text segments are marked as *pass*. We also examined a few instances where the false friends are actually used as a decisive factor in the misalignment detection process, and we found all of the aligned text segments to be true translations. It is often the case that one word of the false friend pair is the translation of another word in the text segment of the other language.

In alignment problem detection, for mismatches in numbers, the system would mark the pair as *problem*. However, in actual fact the texts may have no translation and alignment problems at all. For example, for different phone numbers of contacts for different languages, the discrepancy in numbers does not mean that the sentences are misaligned or the translation is problematic. It is the right way to indicate that, for different languages, different phone numbers should be called. The same is true with the use of numbers indicating a particular year. When we say 2006 in French and say 'this year' in English, it becomes a 'problem' match because of different ways of expressing the same content meaning, but in such cases the lack of the number '2006' in English does not constitute a misalignment. Thus, the label 'problem' becomes problematic in this instance.

5 Conclusion

In this paper, we have described the general framework of SDTES, a system for the automatic extraction of translation pairs from web-based bilingual materials. 70,555 translation segments were extracted from the StatCan *Daily* data bank which was composed of the news release texts from 2004 to 2006 at Statistics Canada. The length-based statistical model and other algorithms were used together with the structural and textual features of the HTML texts for alignment of text segments and for detection of the misaligned parts. The aligned pairs were further filtered and formatted for easy portability into the TM systems, bilingual IR databases and other NLP applications. The evaluation and results show that the automatic translation extraction system is a very suitable model for generating translation pairs from the type of web-based document of which *Daily* texts are an example. Introduction of the HTML structural features as major text block anchoring points prior to the utilization of the Gale-Church statistical method greatly enhanced the performance of the statistical parallel alignment and the capacity of thwarting massive misalignment. The SDTES system demonstrates that textual features such as cognates and punctuation marks can play a very important role in aiding the misalignment detection process. The system introduces a two-step method of identifying cognates in bitexts. The AMS search based on the candidate matching list produced by the K-vec algorithm performed very reliably and accurately. Experiments with bilingual texts from five other government websites have shown that the major components of the system, with some parameters properly tuned, can be adapted for translation extraction of other web-based Canadian

government publications. The extracted translations can be readily fed into TM systems, MT systems, CLIR systems and other NLP systems.

Acknowledgements Thanks go to the managers and staff of Statistics Canada at various levels: the Communications and Information Services Branch, the Communications and Library Services Division, and the Official Languages and Translation Division. They gave us full support for this research project. We are also indebted to Professor Andrew Brook and Professor Jim Davies of Carleton University, to Professor David Sankoff of University of Ottawa, to Dr. Joel Martin of National Research Council Canada, to Dr. Collin Baker of the International Computer Science Institute, to Professor Charles J. Fillmore of University of California at Berkeley and ICSI, and to the anonymous reviewers of this paper, for their valuable comments and suggestions.

References

Brown PF, Lai JC, Mercer RL (1991) Aligning sentences in parallel corpora. In: 29th annual meeting of the association for computational linguistics, Berkeley, CA, pp 89–94

Callison-Burch C, Bannard C, Schroeder J (2005) A compact data structure for searchable translation memories. In: 10th European association for machine translation conference: building applications of machine translation, conference proceedings, pp 59–65

Chen S (1996) Building probabilistic models for natural language. PhD thesis. Harvard University, Cambridge, MA

Chen A, Gey F (2001) Translation term weighting and combining translation resources in cross-language retrieval. The tenth text retrieval conference (TREC 2001), Gaithersburg, MD, pp 529–533

Chen J, Nie JY (2000) Parallel web text mining for cross-language IR. In: Proceedings of RIAO 2000: content-based multimedia information access, vol 1, Paris, France, pp 62–78

Chesñevar C, Sabaté M, Maguitman A (2006) An argument-based decision support system for assessing natural language usage on the basis of the web corpus. Int J Intelligent Syst 21(11):1151–1180

Danielsson P, Mühlenbock K (2000) The misconception of high-frequency words in Scandinavian translation. In: Envisioning machine translation in the information future, 4th conference of the Association for Machine Translation in the Americas, AMTA 2000, Cuernavaca, Mexico, pp 158–168

Deng Y, Kumar S, Byrne W (2006) Segmentation and alignment of parallel text for statistical machine translation. J Nat Lang Eng 12(4):235–260

Fung P, Church KW (1994) K-vec: a new approach for aligning parallel texts. In: Proceedings of the 15th international conference on computational linguistics, Kyoto, Japan, pp 1096–1102

Gale WA, Church KW (1991) A program for aligning sentences in bilingual corpora. In: 29th Annual meeting of the association for computational linguistics, Berkeley, CA, pp 177–184

Gey FC, Chen A, Buckland MK, Larson RR (2002) Translingual vocabulary mappings for multilingual information access. In: 25th Annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2002), Tampere, Finland, pp 455–456

Hutchins J (2005) Towards a definition of example-based machine translation. In: MT summit X. Workshop: second workshop on example-based machine translation. Phuket, Thailand, pp 63–70

Inkpen D, Frunza O, Kondrak G (2005) Automatic identification of cognates and false friends in French and English. In: Proceedings of international conference on recent advances in natural language processing (RANLP'05), Borovets, Bulgaria, pp 251–257

Jutras J-M (2000) An automatic reviser: the transCheck system. In: 6th applied natural language processing conference and 1st meeting of the North American chapter of the association for computational linguistics, proceedings of the conferences, Seattle, WA, pp 127–134

Kay M, Röscheisen M (1993) Text-translation alignment. Comput Linguist 19(1):121–142

Kondrak G, Dorr B (2004) Identification of confusable drug names: a new approach and evaluation methodology. In: Coling, 20th international conference on computational linguistics, proceedings, Geneva, Switzerland, vol II, pp 952–958

Li KW, Yang C (2006) Conceptual analysis of parallel corpus collected from the Web. J Am Soc Inf Sci Technol 57(5):632–644. doi:10.1002/asi.20326

McArthur T (ed) (1992) The Oxford companion to the English language. Oxford University Press, Oxford

Melamed ID (1999) Bitext maps and alignment via pattern recognition. Comput Linguist 25(1):107–130

Moore RC (2002) Fast and accurate sentence alignment of bilingual corpora. In: Machine translation: from research to real users, 5th conference of the Association for Machine Translation in the Americas, AMTA 2002, Tiburon, CA, proceedings, LNAI 2499. Springer, Berlin, pp 135–144

Neumann S, Hansen-Schirra S (2005) The CroCo project. Cross-linguistic corpora for the investigation of explicitation in translations. In: Proceedings from the corpus linguistics conference series, Corpus Linguistics 2005, Birmingham, UK, vol 1, no. 1. Available online: http://www.corpus.bham.ac.uk/PCLC/cl-134-pap.pdf

Pedersen T, Varma N (2002) K-vec++: approach for finding word correspondences. Available online: http://www.d.umn.edu/~tpederse/parallel.html

Resnik P (1999) Mining the web for bilingual text. In: 37th annual meeting of the association for computational linguistics, College Park, MD, pp 527–534

Resnik P, Smith N (2003) The web as a parallel corpus. Comput Linguist 29(3):349–380.

Ribeiro A, Dias G, Lopes G, Mexia J (2001) Cognates alignment. In: MT summit VIII, machine translation in the information age, proceedings, Santiago de Compostela, Spain, pp 287–292

Sánchez-Villamil E, Santos-Antón S, Ortiz-Rojas S, Forcada ML (2006) Evaluation of alignment methods for HTML parallel text. Advances in natural language processing, proceedings of FinTAL 2006, 5th international conference on natural language processing, Turku, Finland, LNCS 4139. Springer, Berlin, pp 280–290

Sigurbjörnsson B, Kamps J, de Rijke, M (2005) EuroGOV: engineering a multilingual web corpus. In: Accessing multilingual information repositories: 6th workshop of the cross-language evaluation forum, CLEF 2005. LNCS 4022. Springer, Berlin, pp 825–836

Simard M, Foster G, Isabelle P (1992) Using cognates to align sentences in bilingual corpora. In: Fourth international conference on theoretical and methodological issues in machine translation, empiricist vs. rationalist methods in MT (TMI 92), proceedings of the conference, Montreal, Canada, pp 67–81

Simões AM, Almeida JJ (2006) Combinatory examples extraction for machine translation. In: 11th annual conference of the European association for machine translation, proceedings, Oslo, Norway, pp 27–32

Tufiş D, Barbu AM, Ion R (2004) Extracting multilingual lexicons from parallel corpora. Comput Hum 38(2):163–189.

Véronis J (2000a) Alignement de corpus multilingues. In: Pierrel J-M (ed) Ingénierie des langues, Traité IC2-Série Informatique et SI. Èditions Hermes Science, Paris

Véronis J (2000b) From the Rosetta stone to the information society: a survey of parallel text processing. In: Véronis J (ed) Parallel text processing: alignment and use of translation corpora. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 1–24

Wang JH, Teng JW, Lu WH, Chien LF (2006) Exploiting the Web as the multilingual corpus for unknown query translation. J Am Soc Inf Sci Technol 57(5):660–670

Wu D (1994) Aligning a parallel English–Chinese corpus statistically with lexical criteria. In: 32nd annual meeting of the association for computational linguistics, proceedings of the conference, Las Cruces, NM, pp 80–87