Language Translation of Web-based Content

Bart Kahler, Brian Bacher, and K.C. Jones
Science Applications International Corporation
3745 Pentagon Way
Beavercreek, OH 45431
kahlerb@saic.com

***Abstract** – Machine Translation (MT) software today provides adequate conversion of foreign languages to one's native tongue; however, dialects, slang, and character conversion errors result in partially successful translations. For an accurate translation, a native speaker is often required to correct the translation by using sentence structure and word use cues to capture the true meaning. MT character conversion from Cyrillic, Asian, and Arabic languages to western characters induce errors in the translated text which can change the meaning or result in characters being associated together that do not form words. The authors present a solution using open source MT and the International Organization for Standardization (ISO) character mapping. The solution provides proper character conversion to achieve greater translation accuracy for web-based content.*

***Keywords**: Machine Translation (MT), Cross-Language Information Retrieval (CLIR) **[1]**, Computer-Aided Translation (CAT), International Organization for Standardization (ISO)*

1 Introduction

A standard character encoding is necessary if communication between computers is to be achieved. Prior to the mid-1960's different types of computers could not digitally exchange information because each type or family of computers used a unique character encoding system. In 1964 the American Standard Code for Information Interchange (ASCII) was released by the American National Standards Institute (ANSI). ASCII used an encoding scheme which pairs a character with a number code, representing a total of 128 characters. The extended ASCII character set represents 256 characters. **[2]**

Although ASCII is capable of representing Roman characters such as those of the English language, languages using non-Roman characters such as Eastern languages are ignored. In the late 1980's work on a universal character encoding system known as Unicode began to address the needs of world languages by using fixed width 16 bit character encoding. The Unicode Consortium now maintains the Unicode Standard and working with the International Organization for Standardization (ISO) developed the Unicode Standard ISO/IEC 10646. **[3,4]**

The most popular character encoding used on the World Wide Web today is Unicode Transformation Format-8 (UTF-8) **[5]**. UTF-8 is an 8-bit encoding standard designed to represent every character in the Unicode Standard and is backward compatible with ASCII. Shown in Figure 1 are the top 10 character encodings used by Websites. A Website may use multiple character encodings. The remaining 1.8% of encodings used by Websites include ISO-8859-15, ISO-8859-9, Windows-1250, EUC-KR, Windows-1254, Big5, Windows-874, US-ASCII, ISO-8859-7, and TIS-620 **[6,7]**.

Web information may be useful, but only when the end user can read the content, preferably in their native language. More than 2.1 billion people across 195 countries are connected to the Internet **[8]**. Users of the Web access over 8.21 billion Websites containing Web content in one of 35 different languages **[6,9]**. The top 10 languages found on the Internet are shown in Figure 2. More than half of Web-based content is in the English language with the remaining Website content divided among the other 34 languages.

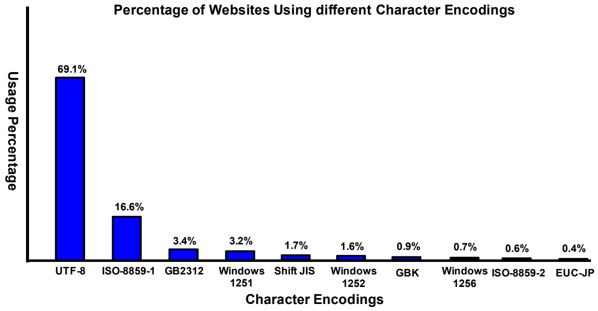

Figure 1. Top 10 character encodings used by Websites **[6,7].**

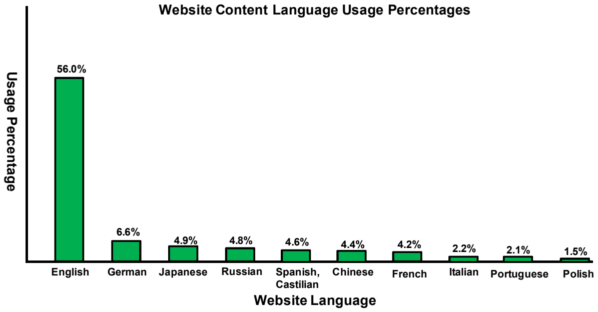

Figure 2. Top 10 Languages Used by Websites **[6,9]**

English is the predominant language used for Web-content today. English is often used as a bridging language when converting between two other languages **[10]**. English is frequently taught as a second language throughout the world. For these reasons our paper will focus on converting non-English Web content into the

978-1-4673-2792-3/12/$31.00 ©2012 IEEE

English language. However; the design techniques presented in this paper can be applied to other machine language translations. In Section 2 the authors present language translation issues and techniques. In Section 3 language translation evaluation and metrics are presented. A machine translation design flow for Web content is discussed in Section 4. Our conclusions are in Section 5. References are found in Section 6.

2 Language Translation

Language translation is a multi-billion dollar business that is growing significantly each year. There are two main translation areas where tools are needed; MT of text **[11,12]** and MT of speech **[13,14]** either into text or into speech of another language. This paper will focus on the MT of text.

2.1 Issues

Understanding the factors that will influence the accuracy and precision of a machine translation (MT) are important. The structure of the language and grammatical standards of the language being translated will impact the resulting translation. Languages have different roots and therefore syntax or sentence structure may differ. Various sentence construction methods may be used with subject-object-verb the most common at 45% followed by subject-verb-object at 42% **[15]**.

Multiple word meanings can be problematic. For instance the word *light* can be a noun, verb, or adjective depending on sentence placement as shown in Table 1. One to one word representation is not always possible because a source language word may be represented by many target language words.

Noun	He turned on the *light.*
Verb	Please *light* the candle.
Adjective	She ate a *light* dinner.

Table 1. Multiple meanings of the word *light*.

Languages contain word usages which can be grammatical and natural for a specific language. Idioms such as '*It is raining cats and dogs.'* can prove difficult to translate properly. The meaning of the entire sentence can't be predicted from the meanings of the individual words **[16]**. Consequently, a direct conversion from the source language to the target language will yield little meaning because the target language has no equivalent translation.

2.2 Techniques

Machine translation may be accomplished via rule based or statistical methods. Rule based approaches were the earliest successful method of MT. However, the digital information age has led to the development of statistical translation techniques which are now widely used.

Rule based MT relies on an understanding of sentence structure for each language and requires a bilingual dictionary **[17,18,19,20]**. Each phrase from a source language is broken down into individual words and labeled as sentence parts. The dictionary converts each word into the new language. The sentence gets reassembled in the new language using the new language's grammar rules. A literal translation of the sentence is produced. Rule based approaches are resource intensive and require linguistic experts to tune the system. Issues such as idioms and multiple word meanings may not be entirely accounted for when developing the system. One of the earliest MT companies is SYSTRAN who successfully exploited rule based MT systems **[17]**.

Statistical machine translation (SMT) relies on common literature translated in multiple languages. A great source for parallel translated text is the official writings of the UN which translates everything into the six official languages (Arabic, Mandarin Chinese, English, French, Russian, and Spanish) **[21]**. SMT takes phrases from parallel translated text and creates a translating dictionary based on the location of the words in a phrase from the source language compared to the target language **[22,23,24]**. Phrase based SMT is the most commonly used method because phrases provide better context of words rather than individual words. Syntax based SMT is being used to improve translation by accounting for parts of the sentence structure. Since SMT does not use grammar rules, rapid development is possible provided parallel translated text is available for the source and target languages.

Hybrid machine translation (HMT) is a technique using statistical methods coupled with rule based approaches in an attempt to improve overall translation performance. Science Applications International Corporation (SAIC) offers HMT technology enabling contextual translations which are adaptive and customizable by the user **[25]**. For example, the vocabulary can be specialized for the medical field to allow doctors and nurses to communicate medical information to patients fluent in a different language **[26]**. SYSTRAN has converted to an HMT to improve fidelity and meaning of the resulting translations produced by their tools.

2.3 Commercial Products

A number of commercial products are now available for MT of Web based content. Three popular search engines which have MT capabilities are Yahoo's Babelfish **[27]**, Google's Google Translate **[28]**, and Microsoft's Microsoft Translator **[29]**. Yahoo's Babelfish uses SYNTRAN's software to generate translations based on a hybrid system using SMT with rule based post processing. Google Translate uses a statistical method with UN documents as its main training data. Microsoft Translate is a linguistically informed SMT model. All three MT systems rely on a statistical model at the core.

Babelfish, Google Translate, and Microsoft Translator were initially open source translation tools. However, language translation is a growing multi-billion dollar business and all three of these Web based translation systems will soon be charging for their services **[30, 31]**. More than 8 billion websites exist and 44% of the content is written in languages other than English. Translating all of the Web content into English will prove to be costly. A search for the term "Iran nuclear weapons" in Arabic (إيران أسلحة نووية)on Google **[32]** in March of 2012 produced 3.18 million results. The cost would be $6,598.50 with the Microsoft Translator and $15,900.00 with Google Translate if ten percent of the results were translated with each site averaging 500 words at 5 characters per word. Below is a chart detailing current translation costs. Yahoo's Babelfish is not shown because SYSTRAN's pricing is not public knowledge.

Translation Service	**Price plan**	**Price per word**	**Price per 500 word Web page**	**Number of Web pages for $1000**
Google Translate	$20.00 per 1 M chars	$0.0001000	$0.05	20,000
Microsoft Translator	$4150.00 per 500 M chars	$0.0000415	$0.02	48,193

Table 2. Cost for using commercial translators over the Internet

An open source statistical machine translation system known as Moses is freely available **[33]**. Moses uses phrased-based and tree-based SMT methods. A collection of parallel translated texts are required to train Moses. An SMT based on a Moses core could be created with some research and developed into a customized translation tool. Such an SMT development approach helps to keep costs down by removing licensing and use fees now being added to other mature open source translation services.

3 Metrics

The evaluation of MT systems is accomplished by comparing translated results to the correct human translation. To assess MT performance, common written text such as stories translated into multiple languages are used. For example, *Little Red Riding Hood* is a children's story that has been translated into many languages and may be used to test MT capabilities. A Spanish source text for part of *Little Red Riding Hood* and the corresponding correct English translation are found in Table 3 **[34]**. Three different English MT results using the Spanish source text are included as part of Table 3.

Description	**Text**
Spanish Source Text **[34]**	"Abuela, ¿por qué tienes los ojos tan grandes?" Caperucita Roja preguntó. "Para que yo pueda ver mejor," Dijo la abuela. "¡Oh, abuelita, ¿por qué tienes la boca tan grande?" "Para poder comerte mejor!" Entonces, la abuela salta de la cama.
Correct English Translation **[34]**	"Grandma, why do you have such big eyes?" Little Red Riding Hood asked. "So that I can see better." the grandma said. "Oh, Grandma, why do you have such a big mouth?" "So I can eat better!" Then, the grandma jumps out of the bed.
Google Translate Result	"Grandma, why are your eyes so big?" Little Red Riding Hood said. "So I can see better," said the grandmother. "Oh, Grandma, why have the big mouth?" "To eat better!" Then the grandmother jumps out of bed.
Microsoft Translator Result	"Grandmother, why have such large eyes?" Little Red Riding Hood asked. "So I can see better," said the grandmother. "Oh, grandmother, why have the big mouth?!" "To be able to eat better!" Then, Grandma jumps out of the bed.
Yahoo Babelfish Result	" Grandmother, why you have the so great eyes? " Red Caperucita asked. " So that I can see better, " The grandmother said. " Oh, grandma, why you have the so great mouth? " " In order to be able comerte better" Then, the grandmother jumps of the bed.

Table 3. Spanish to English Translation Example

Comparison of the three translation software results in Table 3 to the correct English translation reveals differences in each MT system's performance. Google Translate and Microsoft Translator conveyed the basic meaning with Google Translate producing a more readable translation. Yahoo Bablefish didn't translate two words; "Caperucita" and "comerte". The un-translated words are important to the meaning of the story.

The need for a consistent, unbiased, assessment of translation performance is apparent in the small example of Table 3. A method of evaluating single sentences with parallel translations to group them into one of five categories listed in Table 4 was presented in the literature **[24]**. Totaling how many sentences fall in each category is a subjective attempt to put a number on translation performance.

Category	**Description**
Exact	MT translated sentence is exactly the same as the parallel translation
Alternate	MT translated sentence conveys the same meaning but the words are different or in an alternate order as the parallel translation
Different	MT translated sentence is a legitimate translation that does not convey the same meaning as the parallel translation
Wrong	MT translated sentence forms a sentence but cannot be interpreted as the parallel translation
Ungrammatical	MT translated sentence is grammatically deficient

Table 4. Translation Categories [24]

The translations of the Spanish text into English found in Table 3 were assessed using the five categories defined in Table 4. The evaluation technique especially for a large number of sentences is a time consuming process. The MT performance for the example is presented in Table 5.

Category	Google Translate	Microsoft Translator	Yahoo Babelfish
Exact	0	1	0
Alternate	5	3	2
Different	0	0	0
Wrong	0	0	2
Ungrammatical	1	2	2

Table 5. Translator performance for example

A computational MT evaluation approach would be faster and less subjective. The MT community has recognized two computational evaluation techniques; Bilingual Evaluation Understudy (BLEU) **[35,36,37]** and the National Institute of Standards and Technology (NIST) technique **[38,39]**.

BLEU is an objective translation evaluator providing discrete performance scores. A BLEU score evaluates word position in a sentence comparing a human translation to the machine translation. The algorithm determines if the MT has all of the words in the human translation and checks the order of the words. Grammar rules are not applied to verify if a sentence is formed.

The BLEU score **[36-39]** which ranges between 0 and 1 is defined in Equation 1 as

$$BLEU = BP \times \exp(\textstyle\sum_{n=1}^{N} w_n \log p_n) \qquad (1)$$

where $w_n = N^{-1}$, the precision score p_n is given by Equation 2

$$p_n = \frac{\sum_{s \in C} \sum_{n-gram \in S} Count_{matched}(n-gram)}{\sum_{s \in C} \sum_{n-gram \in S} Count_{matched}(n-gram)} \qquad (2)$$

Where $\sum_{s \in C} \sum_{n-gram \in S} Count_{matched}(n - gram)$ is the number of words in the translation that match the reference source translation and $\sum_{s \in C} \sum_{n-gram \in S} Count_{matched}(n - gram)$ is the total number of words in the translation being evaluated. The brevity penalty is determined in Equation 3 by

$$BP = \begin{cases} 1 & if\ c > r \\ e^{1-r/c} & if\ c \le r \end{cases} \qquad (3)$$

where c is the length of the translation and r is the source text length.

The NIST technique is an improved translation evaluator building upon the BLEU method **[38,39]**. NIST uses an arithmetic mean with weights applied based on the informative nature of the words. Less frequently used words have a higher value than more commonly used words. The weights are determined by Equation 4 as

$$Info(w_1 \cdots w_n) = \log_2 \left(\frac{\#\ of\ occurrences\ of\ w_1 \cdots w_{n-1}}{\#\ of\ occurrences\ of\ w_1 \cdots w_n} \right)$$
$$(4)$$

where w_n represents each word in the translation. The NIST score is determined via a modified BLEU algorithm and is shown in Equation 5.

$$NIST\ Score =$$
$$\sum_{n=1}^{N} \left\{ \sum_{\substack{all\ w_1 \cdots w_n \\ that\ co-occur}} Info(w_1 \cdots w_n) \Big/ \sum_{\substack{all\ w_1 \cdots w_n \\ in\ sys\ output}} (1) \right\} \cdot$$
$$exp\left\{ \beta \log^2 \left[\min \left(\frac{L_{sys}}{\overline{L_{ref}}}, 1 \right) \right] \right\} \qquad (5)$$

where β is set to force the brevity penalty factor to 0.5 when the number of words in the system output is $\frac{2}{3}$ of the average number of words in the reference translation, $N = 5$, $\overline{L_{ref}}$ is the average number of words in a reference translation averaged over all reference translations, and L_{sys} is the number of words in the translation being scored.

For a more thorough evaluation, collect parallel translated text related to the subject matter the tool will be expected to translate. Foreign news reports are sources of text often with corresponding English translations. Some potential Web sites to use to obtain parallel translated text are given in Table 6.

Arabic	http://www.aljazeera.net
Mandarin Chinese	http://www.xinhuanet.com
German	http://www.spiegel.de
Russian	http://ria.ru

Table 6. Parallel Translation Sites

4 Design

As discussed in the Introduction, representing characters (numbers, symbols, and letters) with numbers is called character encoding. Multiple languages around the world require different character representations. Fortunately, all characters can be encoded into UTF-8 Unicode. UTF-8 is a variable byte sized encoding scheme that can represent up to 4 bytes or 4,294,967,296 characters and is the most widely used encoding scheme for Web pages. Additional character sets are used on web sites. In the Introduction, Figure 1 shows a list of common character set encoding schemes found on the Internet. Language translation software prefers UTF-8 encoding and text should be converted into UTF-8 prior to translation. The encoding scheme being used may be detected by discovering the Web page's character set or encoding declaration. Most Web pages can easily be converted to UTF-8 using a Python library called Beautiful Soup **[40]**. Addition programming is required for pages with missing character set declarations.

Using publically available software over the Internet, here are the steps necessary to translate a Website from a foreign language into English. Find the Web site and download the HTML content. Remove the unnecessary HTML, script tags, and excess white space. If necessary, capture the text and convert it into UTF-8. The authors have found Python's Beautiful Soup and Urllib libraries capable of capturing and converting the text into UTF-8 encoding. Next, detect the foreign language and send the text with its language type to a translator. Finally, capture and display the English text. Found in Figure 3 is a flow diagram of the language translation steps. The process can be programmed in a number of software languages.

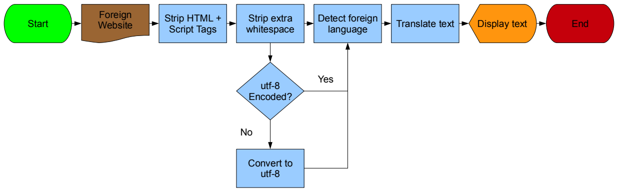

Figure 3. Web-based Language Translation Design Flow

5 Conclusion

Character encoding is important to language translation. We identified the most common character encoding schemes (UTF-8) and the top languages (English) found on the Internet. Typical language translation issues such as idioms must be considered in the design of an MT system. The three machine language translation techniques are Rule based, Statistical, and Hybrid. The cost is rising and demand is increasing for language translation services. The metrics to assess how well a translator is functioning are subjective human evaluation and two computational scoring methods, BLEU and NIST. Creating a system to translate Web-based content can be performed with publically available tools over the Internet. After the desired Website is identified, translation requires removing the HTML tags, preserving the character sets, and sending the text to a language translation tool. Converting a Website to English can be costly. The design of a system to collect information from a foreign Website should use the best translation software for the selected foreign language to ensure the highest quality translation product. Tailoring a MT tool for the specific content to be translated such as medical terms or engineering will improve translation accuracy for the specific topic areas.

The authors propose the creation of an SMT engine using Moses and subject specific text. The translation performance may then be assessed using human evaluation and computational scoring via BLEU and NIST. A well designed machine translation program can both detect and translate a foreign language into English.

6 References

[1] [M. Aljlayl, O. Frieder, and D. Grossman, "On Arabic-English Cross-Language Information Retrieval: A Machine Translation Approach", *IEEE Computer Society, Proceedings of the International Conference on Information Technology: Coding and Computing*, 2002.

[2] M. Brandel, " 1963: The Debut of ASCII", 6 July 1999 CNN - http://edition.cnn.com/TECH/computing/9907/06/1963.idg/index.html

[3] Unicode: Summary Narrative - http://www.unicode.org/history/summary.html

[4] Unicode v. 6.0.0 - http://www.unicode.org/versions/Unicode6.0.0/

[5] Usage of Character encodings for Websites - http://w3techs.com/technologies/overview/character_encoding/all

[6] W3 Techs - http://w3techs.com/technologies/overview/content_language/all , March 2012.

[7] http://www.techterms.com/definition/characterencoding

[8] CIA World Factbook - https://www.cia.gov/library/publications/the-world-factbook/index.html

[9] World Wide Web Size - http://www.worldwidewebsize.com/ , March 2012.

[10] S. Bakhshaei, S. Khadivi, N. Riahi, "Farsi-German Statistical Machine Translation Through Bridge Language", *IEEE 5th International Symposium on Telecommunications*, pp. 557-561, 2010.

[11] C. Yang, "Cross-Language Instant Messaging With Automatic Translation", *IEEE Computer Society, Fourth International Conference on Ubi-Media Computing*, pp. 222-226, 2011.

[12] H. Yu, F. Ren, D. Huang, and L. Li, "Designing Effective Web Minning-Based Techniques for OOV Translation", *IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp. 1-8, 2010.

[13] K. Macherey, O. Bender, and H. Ney, "Applications Of Statistical Machine Translation Approaches to Spoken Language Understanding", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 4, pp. 803-818, May 2009.

[14] H. Ney, S. Nieben, F. Josef Och, H. Sawaf, C. Tillmann, and S. Vogel, "Algorithms for Statistical Translation of Spoken Language", *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 1, January 2000.

[15] R. Tomlin, "Basic Word Order: Functional Principles", Croom Helm, London, pp 22, 1986.

[16] Collins English Dictionary – Complete and Unabridged, HarperCollins Publishers 2003.

[17] M. Osborne, "MT:History and Rule-based system", http://www.inf.ed.ac.uk/teaching/courses/mt/lectures/history.pdf , 2012.

[18] F. Bond, "Machine Translation Introduction", ACL/HCSNet NLP/IR 2006, http://www.csse.unimelb.edu.au/research/lt/nlp06/materials/Bond/mt-intro.pdf , 2006.

[19] D. Demner-Fushman and D. Oard, "The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval", *IEEE Computer Society, Proceedings of the 36th Hawaii International Conference on System Sciences*, 2002.

[20] J. Tenni, A. Lehtola, C. Bounsaythip, and K. Jaaranen, "Machine Learning Of Language Translation Rules", *1999 IEEE International Conference On Systems, Man, and Cybernetics*, Vol. 5, pp. 171-177, 1999.

[21] A. Rafalovitch and R. Dale, "United Nations General Assembly Resolutions: A Six-Language Parallel Corpus", MT Summit, 2009.

[22] Introduction to Statistical Machine Translation - http://michaelnielsen.org/blog/introduction-to-statistical-machine-translation/

[23] W. Wang, A. Stolcke, J. Zheng, “Reranking Machine Translation Hypotheses with Structured and Web-based Language Models”, *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 159-164, 2007.

[24] P. F. Brown, J. Cocke, et al. “A Statistical Approach to Machine Translation”, *Computational Linguistics*, Vol. 16, No. 2, pp. 79-85, June 1990.

[25] D. F. Carr, “SAIC Takes on Google With Speech Translation Apps”, http://www.informationweek.com/thebrainyard/news/workgrouping_team_collaboration_workspaces/231602149, September 2011.

[26] “New integrated offering enables tailored translations in context, enhancing efficiency”, http://investors.saic.com/phoenix.zhtml?c=193857&p=irol-newsArticle&ID=1594084, August 2011.

[27] Yahoo Babelfish - http://babelfish.yahoo.com/

[28] Google Translate - http://translate.google.com/

[29] Microsoft Translator - http://www.microsofttranslator.com/

[30] Google Translate API v2 - https://code.google.com/apis/language/translate/overview.html

[31] Microsoft Translation API - http://www.microsofttranslator.com/dev/

[32] Google – http://www.google.com

[33] Moses Project - http://www.statmt.org/moses/

[34] C. Packer, “Translation Software: Which Are the Most Accurate?”, http://translation-software-review.toptenreviews.com/translation-software-which-are-the-most-accurate.html

[35] http://translation-software-review.toptenreviews.com/the-translation-test-format.html

[36] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “BLUE: a Method for Automatic Evaluation of Machine Translation”, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311-318, 2002.

[37] C. Callison-Burch, M. Osborne, and P. Koehn, “Re-evaluating the Role of BLEU in Machine Translation Research”, *11th Conference of the European Chapter of the Association for Computational Linguistics: ECAL 2006*, pp. 249-256, 2006.

[38] D. Ze-ya, Z. Han-fen, Z. Quan, M. Jian-ming, and C. Yu-huan, “Automatic Machine Translation Evaluation Based on Sentence Structure Information”, *IEEE Computer Society, 2009 International Conference on Asian Language Processing*, pp. 162-166, 2009.

[39] G. Doddington, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”, *Proceedings of the Second International Conference On Human Language Technology Research*, pp. 128-132, 2002.

[40] http://www.crummy.com/software/BeautifulSoup/