Source-Target Mapping Model of Streaming Data Flow for Machine Translation

Jolanta Mizera-Pietraszko
Faculty of Mathematics, Physics and Computer Science
Opole University
Opole, Poland
jmizera@math.uni.opole.pl

Grzegorz Kołaczek
Faculty of Computer Science and Management
Wroclaw University of Science and Technology
Wroclaw, Poland
Grzegorz.Kolaczek@pwr.edu.pl

Ricardo Rodriguez Jorge
Department of Industrial and Manufacturing Engineering
Autonomous University of Ciudad Juarez
Chihuahua, Mexico
ricardo.jorge@uacj.mx

Abstract — **Streaming information flow allows identification of linguistic similarities between language pairs in real time as it relies on pattern recognition of grammar rules, semantics and pronunciation especially when analyzing so called international terms, syntax of the language family as well as tenses transitivity between the languages. Overall, it provides a backbone translation knowledge for building automatic translation system that facilitates processing any of various abstract entities which combine to specify underlying phonological, morphological, semantic and syntactic properties of linguistic forms and that act as the targets of linguistic rules and operations in a source language following professional human translator. Streaming data flow is a process of mining source data into target language transformation during which any inference impedes the system effectiveness by producing incorrect translation. We address a research problem of exploring streaming data from source-target parallels for detection of linguistic similarities between languages originated from different groups.**

Keywords— streaming data flow; pattern recognition; data mining; machine translation; natural language processing;

I. Introduction

Translation knowledge has been widely studied since the time of developing first machine translation systems with the aim at extracting language-pair phenomena from linguistic ontology resources like dictionaries, thesauruses, and non-ontology ones like parallel and comparable corpora from the one hand, and adopting the language rules based on the knowledge gained about a particular language to build infrastructure for increasing expectations on automatic translation reliability from the other hand.

Morphological transfer model built by Melcuk and Wanner [1] improves processing tenses transitivity while translating Russian and German. McCary et al. [2] propose acquiring data from parallel texts and Tree banks for Arabic and Chinese into English translation. Knowledge extraction for translation patterns from sentence-leveled bilingual corpora is presented by McTait [3]. Transliteration while translating Bangla into English resulted in 8 to 9% accuracy by BLUE measure [4]. Zhai [5] adopt word association mining in language modeling to translation from Arabic into English language pair. For identifying the so-called false friends and cognates from comparable corpora Mitkov et al. [6] discuss a mapping method in English, French, German and Spanish with the aim at translating the language pairs. Hung [7] has developed multilingual machine translation system for identifying source language and coding it by segmentation of the source texts. Another work by Wu and Wang [8] presents a pivot language as a tool for automatic translation of two other languages with their identified phenomena. The literature study presented above provides an evidence for extensive usage of knowledge acquired from the analysis of streaming data flow for improving effectiveness of automatic translation.

II. Research Problem & Methodology

This paper addresses a research problem of improving automatic translation quality by applying knowledge about source-target mapping to identify language-pair phenomena between Polish, English and French. In communication theory by Shannon E. C. [9], an information source produces a message to be sent to a transmitter that transforms it into a signal, or a sequence of signals for transmission over the channel. Then, the receiver reconstructs the message from the signal while delivering it to the receiver. Applying this rule of communication theory to automatic translation process, we define a source text $S=\{s_1, s_2, \cdots, s_n\}$ as an information source that is to be sent in real time to the transmitter being a machine translation system (MTS), in particular a translation model, which transforms it then, into a signal for a receiver as the system output - a target text $T=\{t_1, t_2, \cdots, t_n\}$ covering some meaning which represents a message. Index n denotes a number of the message segments in both S and T constituting the word classes which belong to the same linguistic category.

$$S \rightarrow MTS \rightarrow T \qquad (1)$$

When T is a constant sequence of word classes aligned with S in such a way that every word class $t_i \in T$ of set T is uniquely paired with $s_i \in S$ of set S, it is processed backwards by the MTS generating a set of word classes

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

having been translated that is denoted by *S'={s'₁, s'₂, ... , s'ₙ}* on a comparable basis *O={*o₁*, o₂, · · · , oₙ}*, where *oᵢ={*0*,* 1*}* stands for binary translation output, in which 1 denotes that the particular word class and its equivalent, represent the same word class string, while 0 denotes that the translation of the word class is totally incorrect. The coverage $S \equiv S'$ of the word classes translated correctly is reflected in the binary sequence that depicts an encoded message *O*.

$$T \rightarrow MTS \rightarrow S' \Rightarrow S \equiv S'$$

$$o_i = \begin{cases} 1 \; for \; s_i' = s_i \\ 0 \; otherwise \end{cases} \tag{2}$$

Again, dependences between variables *S, S'* and *T* follow the Noisy Channel rule according to which, communication over the channel would not be possible otherwise [10]. Whether the translation is correct or not, it is still possible to get the message in a target language, even despite the poorer information quality. Prospective noise, if any, is produced by a machine translation system (MTS), which does not support an appropriate language resources for processing a particular word class. Alternatively, it may be an input unsupported by the MTS, which results with the target text that is unidentified in the source language.

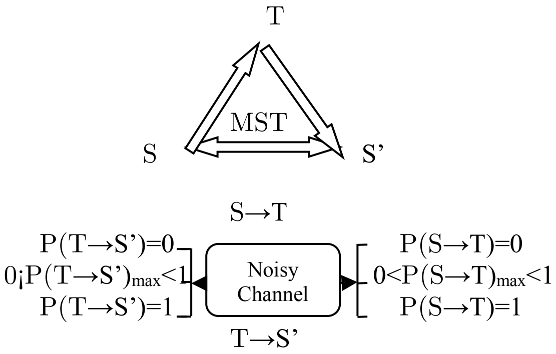

Fig.1 Source-Target Mapping Model

Fig. 1 shows the streaming data flow in our Source-Target Mapping (STM) model. In the first phase, source text part is submitted for translation to produce the target equivalent. Then, following is a backward translation based on the assumed reversibility of the translation algorithm, as symmetric relation, simply saying

$$\bigvee_{i=1}^{N} \{a_i, b_i \in Sym: if \; (a_i)Sym(b_i) \; then \; (b_i)Sym(a_i) \tag{3}$$

Communication theory proves that three alternatives should be considered at each of these two phases: *P(S→T)=*1 probability of source-target mapping or their asymmetrical backward translation *P(T→S')=*1 equals 1 meaning it is perfect, *0<P(S→T)ₘₐₓ<*1 less than perfect in which case the system selects the translation with the maximal value (the most similar to the reference translation in which mapping is a complete coverage of them both) out of all the candidate translations computed by the algorithm and finally *P(S→T)=*0 which denotes that there is no translation for this input text. One of the reasons behind it may be incorrect input text. The last phase relies on the comparison between *S* and *S'* which allows identification of the language pair phenomena as well as the translation algorithm including its model.

Lemma 1.

$$\bigvee_{i=1}^{N} \{S_i \equiv S_i' \in T \,|\, S_i \rightarrow T\} \tag{4}$$

For any source text *S* equal to its asymmetric backward translation equivalent, the target text *T* is a correct translation of the source text *S*.

Proof: Let us apply transitive law that is

$$\bigvee_{i=1}^{N} \{a_i, b_{i,} c_i \in Tr: if \; (a_i)Tr(b_i) \; and (b_i)Tr(c_i) \; then \; (a_i)Tr(c_i)\} \tag{5}$$

Let us denote *P(X)=1* as the correct translation of text *X* (here *S, S'* and *T*) and then let us denote the same translation process as relation *Tr(x).* Since *S* is a reference text, then *P(S)=*1. On the other hand, we assume in (4) that $S_i \equiv S_i'$ such that *P(S')* being in the same relation *Tr(s')* with *S* and *T* meaning *Tr(s)* and *Tr(t)*| where {*s, s', t*∈ *Tr*} owning to the fact that it is the same system equals *1.* Consequently, this implies that transitive relation *Tr* is only on *P(T)=*1. ■

Lemma 2.

$$\bigvee_{i=1}^{N} \{s_i \equiv s_i' \in N \,|\, s_i \neq s_i': t_i \notin T\} \tag{6}$$

For every text segment belonging to set *N*, however different from its equivalent s_i', the matching text segment is an incorrect translation of s_i.

Proof: is the same as above with the exception that the text the segments are treated as the text. Also, it is possible to denote the text segments as the elements of vectors *S, S'* and *T* respectively. ■

III. Bayes Inference Rule

To simplify the research concept let us assume that we have a Bayes network which semantics relies on the comparative bidirectional arrow symbolizing two variants of the network. As long as the texts are comparable we can also agree that for the purpose of the analysis that its one-dimensional relation *S→S'* is satisfactory. Our translation system efficiency is supposed to be at 50% only. Depending upon the efficiency level that is whether the first phase is assessed as correct or not, the system processes or not, the asymmetric backward translation. Let us try to estimate a chance of reaching the phase of equivalence $S_i \equiv S_i'$,

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

P(S)=(T=0.5; F*=0.5)* where *T** and *F** denotes True and False while *P(S, T, S')=P(S|T,S') × P(T|S') × P(S')*

S	*T**	*F**
	0.7	0.3

S-T	*T**	*F**
*T**	1	0
*F**	0	1

S□

S	*T*	*T**	*F**
*T**	*T**	1	0
*T**	*F**	0.5	0.5
*F**	*T**	0.3	0.7
*F**	*F**	0	1

$$P(S' = T^*|S = T^*) = \frac{P(S = T^*, S' = T^*)}{P(S = T^*)} = \frac{\sum_{S\in\{T*,F*\}} P(S = T*, T, S' = T)}{\sum_{S,T\in\{T*,F*\}} P(S = T*, T, S')} \tag{7}$$

Given the correct source text, the chance of getting the correct asymmetrical backward translation for the 50% efficient system equals 1, which indicates that for each system where it is true that *P(S→T)*=1 and *P(S)*=1 as the reference text, *S'* is identical as *S* such that the mapping is complete $S_i \equiv S_i'$. The rule works well even for not very precise translation system as the one analyzed. However, when all the variables in the contingency tables of an a-posteriori probability are changed to 0.5, the same result will drop to 0.5 only, in which case *S≠S'* – this proves the truth of Lemma 1 and 2.

IV. Information Noise in Processing Translation

Our source text has been divided into 9 segments *S=(s₁, s₂, s₃,…s₉)*, each of which is to be translated separately. After the translation, text *T* has been deformed. Asymmetrical backward translation pointed out that segments 4, 5, 7 and 8 are incorrect in texts *T* and *S'*. Assuming that the language competence does not allow the user to assess the translation quality *S→T*, it is enough to evaluate the phase *T→S'* for agreeing the set probability of processing *T→S*, or alternatively by analogy, probability of the translation loss. Let us denote *before* and *after* as the target text being the source text of the phase *T→S'* before and after translation into text *S'*, respectively. $T^{(*)}$ represents time while + and – are correct and incorrect translation quality *Q*. We create contingency tables of an a-posteriori probability.

$T^{(*)}$- *Q*	+	-	$\sum(+,-)$
before	5	4	9
after	5	4	9
$\sum(before, after)$	10	8	18

$T^{(*)}$- *Q*	1	0	$\sum(1,0)$
1	5/18	4/18	9/18
0	5/18	4/18	9/18
$\sum(1,0)$	10/18	8/18	1

P(T) denotes that the text is correct before entering the translation process, thus *P(T)=P(*$T^{(*)}$=1, *Q*=1*)*=5/18. *P(S')* denotes that the text after the translation has not been deformed, such that *P(S')=P(*$T^{(*)}$=0, *Q*=1)=5/18. A-posteriori probability that the event that the translation quality *T→S'* was correct after the translation equals *P(*$T^{(*)}$=0, *Q*=1)=5/9, while a-posteriori probability that the event that the translation quality of the preceding phase *S→T* was correct equals *P(T)=P(*$T^{(*)}$=1, *Q*=1)=5/9. Such a result implies that the difference between translation quality *Q before* and *after* the translation process equals 0 because text *T* was incorrect before it had been submitted for translation, whereas existing information noise has not caused the deformation to be doubled during this phase. Detecting incorrect text segments follows on the comparison basis of texts *S* and *S'*.

V. Knowledge acquired from the STM Model

In this section presented are two case studies with the aim at verifying our STM model. Each study is based on a different translation system: for the purpose of this work the first one is commercial produced by a Polish company while the other one is an online Bing Translator launched in 2012.

A. Case Study 1

Our Source-Target Mapping (STM) Model for streaming data flow was verified also on a commercial MT system produced by Techland – a leading Polish developer of video games. The research objective was to learn the system efficiency in order to indicate what should be improved by the developer. The system processes Polish and French translation. We agreed to evaluate a linguistic phenomenon such a syntax for questions in the Future Simple tense.

Let us enter a source text *S*="Kiedy pójdziesz do domu?" into the STM which gives an output of the translation *T*=" Quand vas tu aller à la maison?".

S is divided into the following four segments:

- s_1="Kiedy", s_2="pójdziesz" s_3="do" s_4="domu"

While T is divided respectively into the segments:

- t_1="Quand", t_2="vas tu aller", t_3= "à", t_4=" la maison"

Here the *S'*=*S* which gives the perfect coverage.

The number of segments for source and target translations is always equal, since otherwise, mapping cannot be complete in the best case.

Knowledge gained about this particular MT system based on the example question sentence only:

- translation of questions with the syntax like this one in the Future Simple tense is correct
- question words are translated accurately
- the full question structures are recognized
- the system recognizes capital letters usage
- it identifies prepositions of place like “do domu”
- this system recognizes personal pronoun $s_2 \rightarrow t_2$
- it relies on Direct Translation Model supported with ontology resources of the general nature

B. *Case Study 2*

Launched in 2012, Bing Translation services have been transferred from the former Yahoo Babel Fish of 1997 [11]. Multilingual features cover translation of thirteen languages making 169 linguistic pairs altogether.

This time we evaluated the system efficiency in English and French for Lexis on the Idioms category.

Our source text entered was *S*=”He worked himself to the bones” divided into the following segments:

- s_1=”He”, s_2=”worked” s_3=”himself” s_4=”to” s_5=”the bone”

The output *T*=”il c’est travaillé à l’os” divided as follows:

- t_1=”il”, t_2=”c’est”, t_3=“travaillé”, t_4=” à” t_5=”l’os”

As a result of asymmetrical backward translation text S’=”it was worked with the bone” which we divide into:

- s_1=”it”, s_2=”was” s_3=”worked” s_4=”with” s_5=”the bone”

Evaluation of the system based on this one idiom only using our STM model provides us with the following remarks:

- comparative black box analysis of $S \neq S'$ gives the results that the system is not supported with any dictionary of idioms
- capital letters are not recognized
- personal pronoun *he* is translated into *il* which refers to either *he* or *she* causing disambiguation – this result indicates a particular phenomenon between English and French languages
- Simple Past tense is translated precisely with reference to the tenses transitivity to Passé Composé
- Bing Translation does not support reflexive pronounces as well as reciprocal pronounces *himself→c’est→was*, this critical point requires improvement
- Prepositions *to→ à → with* cannot be mapped causing another disambiguation meaning that phrasal verbs are not recognized
- This system is grounded upon direct translation model supported with ontology resources of a general nature, it does not recognize technical language owning to the lack of professional linguistic resources
- Algorithm of language processing relies on statistical methods of machine translation

Our source-target mapping model can be exploited in this framework for any language pairs and for evaluation of any translation model or system. From this perspective it can be found universal.

VI. Approaches to Machine Translation

This section discusses the concepts of the main approaches to machine translation such as statistical (SMT), neural (NMT) and called knowledge-based (KMT). In addition to the presentation of our model framework and its implementation areas, out of all the approaches to machine translation, we propose an experiment aimed at identification of particular language pair phenomena using the systems that exploit the approaches described in this section.

A. *Statistical Machine Translation (SMT)*

This is found one of the most commonly used techniques presented in 1999 by Kevin Knight at Johns Hopkins University of Baltimore, USA [12] during the Summer Language Engineering Workshop on Text Normalization Project Pointers when he announced that Canadian government would produce French-English Hanzard data in the form of a parallel text as they plan to develop a baseline statistical MT system to be distributed to all the research community, which explores morphology, syntax and is supported with ontology resources. The major concept of SMT is simply to find the most likely translation of the given text *S.*

$$\bigvee_{i=1}^{N}\{s_i \in S| \bigwedge_{j=1}^{N} t^{max} = arg\max_{t_j} P(t_j \ |s_i): t_j \in T\} \quad (8)$$

Function *argmax* is said to be the target sentence “out of all such sentences, which yields the highest value for” *P(t | s)* [11]. As $i \neq j$ the number of source sentences may be lower than the number of their translation equivalents. Computation of conditional probabilities are processed based on *n*-grams which are unigrams (usually the sentence words), bigrams, or trigrams. When modeled are unigrams, the words are independent from each other creating a *bag of words* which follows the formula

$$P(u_1, u_2, \dots u_n) = P(u_1) \times P(u_2) \dots P(u_n) \quad (9)$$

For bigrams the formula is

$$P(b_1, b_2, \dots b_n) = P(b_1) \times P(b_2|b_1) \dots P(b_n|b_{n-1} \quad (10)$$

And for trigrams

$$P(t_1, t_2, \dots t_n) = P(t_1) \times P(t_2|t_1) \times P(t_3|t_2 t_1) \dots P(b_n|b_{n-1} b_{n-2}) \quad (11)$$

For *n>3* modeling languages seems counterproductive.

B. *Neural Machine Translation (NMT)*

In November 2016 Google announced launching of neural network MT to its Google Translate services which reduced the errors by 60% [13]. Perhaps this is a reason for which recently, the new technology NMT became a hot topic in research IT community. It relies

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

on deep learning making use of the fact that Google, as the only search engine, indexes all the information resulting from the users' activity all over the world. The text entered is translated in real time making the system environment dynamic. GNMT is still supported by SMT used by Google Translate for years. The groundbreaking functionalities are that the modeling languages, especially those rich morphologically including open vocabulary, has improved greatly, parallel texts can train sentence-to-sentence models, the large models accelerate translation inference which is found the most crucial point and finally normalization length caused the GNMT works on real data.

C. Knowledge-Based Machine Translation (KMT)

KMN relies on ontology resources like dictionaries, thesauri, encyclopedia. This approach comprises mainly Direct MT in which source text undergoes morphological analysis supported with ontology resources, after which word reordering follows to be in concordance with the target language syntax. Another model is transfer-based which differs by replacing word reordering to conversion of syntax from source to target language. The third model developed in 1975 was Interlingua – Latin-based language independent as a reference. Morphological analysis component was replaced to morpho-syntactic analysis.

VII. Framework of the Experiment

To assess the effectiveness of our methodology, analyzed were 50 sentences in Polish, English and French with translation technology relying on Systran Statistical Machine Translation (SMT), Google Translate that exploits Neural Machine Translation model (NMT), and Reference.com which reproduces content from external linguistic ontology resources like dictionaries and thesauruses, called Knowledge-Based Machine Translation (KMT).

TABLE I. Evaluation of detected linguistic pair phenomena

Linguistic phenomena	**EN→PL**			**PL→FR**			**EN→FR**		
	SMT	***NMT***	***KMT***	***SMT***	***NMT***	***KMT***	***SMT***	***NMT***	***KMT***
Capital Letters	0.86	1.00	0.68	0.69	0.45	0.98	0.87	0.98	0.56
Prepositions	0.79	0.87	1.00	0.68	0.76	0.67	0.89	1.00	0.57
Question Words	1.00	0.88	0.97	0.84	1.00	0.78	0.65	0.88	0.74

Table 1 shows the mean values of the identified language-pair phenomena (capital letters, prepositions and question words) in the sentences translated by each of the three models. The number of o_i=1 out of 50, that denotes language-pair phenomena identified, was 32 for PL-EN, 46 for EN-FR and only 26 for PL-FR language pairs. The results enable to use the knowledge about these three language-pair phenomena to evaluate, test and train machine translation systems based on statistical, rule-based and ontology knowledge-based translation models.

VIII. Conclusion

Machine translation processing, regardless of the model type, by analogy, is referred to Noisy Channel rules with the aim to extract pattern-based knowledge between the languages originating both from the same and from different linguistic families. In the study, the emphasis is put on the details of source-target mapping of streaming data flow to detect the language-pair phenomena for evaluating translation models. In particular, it explains irreversibility as opposed to assumed reversibility of the translation and identifies the critical points in processing the Polish language when paired with English and French.

References

[1] I. Melcuk, L. Wanner, "Morphological Mismatches in Machine Translation", Machine Translation, vol. 22, 2008.

[2] S. Strassel, C. Christianson, J. McCary, W. Standerman, Data Acquisition and Linguistic Resources", Handbook of Natural Language Processing, Springer, 2011.

[3] K. McTait, "Translation Patterns, Linguistic Knowledge and Complexity in an Approach to EBMT", J. Name Stand. Abbrev., in press.

[4] M. Roy, F. Popowich, "Phrase-Based SMT for Low Density Language Pair", In: Artificial Intelligence, LNCS, vol. 6085, Springer-Verlag, 2010.

[5] Ch. Zhai, Laveraging Comparable Corpora for CLIR in Resource-Lean Language Pair. Information Retrieval, vol. 16, Springer-Verlag, 2010.

[6] R. Mitkov, V. Pekar, D. Blagoer, A. Mulloni, "Methods of Extracting and Classifying Pair of Cognates and False Friends, Machine Translation, vol. 21, Springer, 2007.

[7] V.T. Hung, Reuse of Free Online MT Engines to Develop a MT System of Multilingual Machine Translation, In: Advances in Natural Language Processing, LNCS, Spinger-Verlag, vol. 3230, 2004.

[8] W. Hua, H. Wang, "Pivot Language Approach for Phrase-Based SMT, Machine Translation, vol. 21, Springer-Verlag, 2007.

[9] C.E. Shannon, "A Mathematical Theory of Communication", The Bell Journal, vol. 27, pp. 529-551, 1948.

[10] D.J.C. MacKey, Noisy Channel Coding: Information Theory, Inference and Learning Algorithms, Cambridge University, Press, UK, 2003

[11] J. Callaham, Yahoo's Babelfish replaced with Google Translator, Neowin LLC, NetShelter Technology Media, USA, 2012.

[12] K. Knight,"A Statistical MT Tutorial Workbook", Summer Language Engineering Workshop on Text Normalization Project Pointers, Johns Hopkins University of Baltimore, USA, 1999.

[13] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnicki, O. Vinyals, G. Corrado, M. Hughes, J. Dean, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", Cornell University Library, USA, pp.1-23, 2016.