Published in: Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Goggi, Sara; Grobelnik, Marko; Maegaard, Bente et al. (Hg.) (2016): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), pp. 4458-4462.

Privacy Issues in Online Machine Translation Services – European Perspective.

Pawel Kamocki, Jim O'Regan
IDS Mannheim / Paris Descartes / WWU Münster
Centre for Language and Communication Studies, Trinity College Dublin
Email: kamocki@ids-mannheim.de, jaoregan@tcd.ie

Abstract

In order to develop its full potential, global communication needs linguistic support systems such as Machine Translation (MT). In the past decade, free online MT tools have become available to the general public, and the quality of their output is increasing. However, the use of such tools may entail various legal implications, especially as far as processing of personal data is concerned. This is even more evident if we take into account that their business model is largely based on providing translation in exchange for data, which can subsequently be used to improve the translation model, but also for commercial purposes. The purpose of this paper is to examine how free online MT tools fit in the European data protection framework, harmonised by the EU Data Protection Directive. The perspectives of both the user and the MT service provider are taken into account.

Keywords: machine translation, privacy, personal data

1. Introduction

During the last couple of decades, English has established itself as the global language, especially on the Internet (it is estimated that 53.7% of all websites use English; Russian comes second with only 6,3%[1]). However, only 25.9% of Internet users speak English as their native language (as of November 30, 2015)[2] Therefore, it has been noted that in order to develop its full potential, global linguistic communication on the Internet requires linguistic support systems (Cribb, 2000), such as Machine Translation (MT).

1.1. MT in Context

MT (or automatic translation) can be defined as a process in which software is used to translate text (or speech) from one natural language to another. This section will briefly present the history of MT and various technologies used in the process.

1.1.1. History

The idea to mechanize the translation process can be traced back to the seventeenth century (Hutchins, 1986); however, the field of machine translation is usually considered to have begun shortly after the invention of the digital computer (Koehn, 2010). Shortly after the WWII, Warren Weaver, a researcher at the Rockeffeler Foundation, published a memorandum named "Translation" in which he put forward the idea to use computers for translation, proposing the use of Claude Shannon's work on Information Theory to treat translation as a code-breaking problem (Hutchins, 1999).

Research and commercial development in Machine Translation continued in the "rule-based" paradigm, in which a dictionary, a set of grammatical rules, and varying degrees of linguistic annotation are used to produce a translation, until the early 1990s, when a group of IBM researchers developed the first "Statistical Machine Translation" system, Candide (Berger et al., 1994). Building on earlier successes in Automatic Speech Recognition, which applied Shannon's Information Theory, the group applied similar techniques to the task of French-English translation. In place of dictionaries and rules, statistical MT uses word alignments learned from a corpus (Brown et al., 1993): given a set of sentences that are translations of each other, translations of words are learned based on their co-occurrence (*the translation model*); of the possible translations, the most likely is chosen, based on context (*the language model*).

1.1.2. Technology and challenges

Machine Translation is used for two primary purposes: assimilation (to get the gist of text in a foreign language), and dissemination (as an input to publication, typically post-edited by translators). Free online services, such as Google Translate, are usually intended for assimilation; the translation services in use at the EU, for dissemination. Consequently, systems for assimilation may trade accuracy for broader coverage, and vice versa.

The prerequisite for building statistical MT systems is the existence of human-translated bilingual (or multilingual) corpora – and the bigger the better. An obvious source of professionally translated multilingual corpora are international organizations such as the United Nations or the European Union, generating a substantial amount of

1 According to: http://w3techs.com/technologies/overview/content_language/all , last accessed March 7, 2016.

2 According to: http://www.internetworldstats.com/stats7.htm , last accessed March 7, 2016.

freely available, high-quality multilingual documents (in 24 languages for the EU and in 6 languages for the UN).
Compared to rule-based MT systems, statistical MT systems are cheaper (at least for widely-spoken languages) and more flexible (a statistical system is not designed specifically for one language pair, but can accommodate to any language pair for which a corpus is available). Also, because statistical MT systems are based on human-translated texts, the output of statistical MT is (or at least can be) more natural, and it naturally adapts well to exceptions (if the corpus contains the phrase, it is effectively not an exception).
Zipf's law states that in a given corpus, the frequency of a word is inversely proportional to its frequency rank: the most frequent word will occur (approximately) twice as often as the second, three times as often as the third, and so on. Conversely, the majority of words (40-60%) are hapax legomena (words which only occur once). As statistical MT is corpus-based, it therefore suffers from the problem of data sparsity due to the high proportion of hapax legomena: longer phrase matches are absent from the translation model; contextual information is absent from the language model, affecting the quality ("fluency") of the output.
Data sparsity is the biggest problem in statistical MT. Although there have been attempts to solve it by using linguistic information, dating back to Candide, the most common approach is to simply add more data. A large amount of websites are available in multiple languages, so crawling the web for parallel text is a common method of collecting corpora (Smith et al., 2013), particularly for the providers of free online MT, such as Google and Microsoft, who also operate search engines and therefore already have access to such data. The use of such data, however, has its own problems, as such documents are often not just translated, but localized: different units of measurement, currency, and even country names (Quince, 2014), because of their collocation, become "translations".
Finally, the quality of MT output depends on the quality of the input. Even the most banal imperfections such as misspellings or grammar mistakes – not uncommon in electronic communications – even if they are barely noticeable to a human translator, can compromise the most elaborate MT systems.

2. Data processing in 'free' online MT services

'Free' online MT services allow users to translate texts of different length: from single words and phrases to multiple paragraphs. These texts can be of various types, including private and professional correspondence, blog entries, social media content, newspaper articles… It is therefore not astonishing that these texts may contain information that is sensitive from the point of view of privacy, and more specifically, constitute personal data. If we take into account the fact that MT is an integral part of such privacy-sensitive services as Gmail or Facebook, this becomes even more obvious.
The concept of personal data is defined in art. 2(a) of the Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data (hereinafter: the Directive). According to this article, personal data shall mean *'any information relating to an identified or identifiable natural person'*. This definition has been further analysed by the Article 29 Data Protection Working Party (hereinafter: WP29) in its Opinion 4/2007, which advocates for a broad understanding of the concept. In particular, according to WP29's analysis it covers not only *'objective'* information (i.e. facts), but also *'subjective'* information (i.e. opinions and assessments). Furthermore, the person that the data relate to (i.e. data subject) can be identified (directly or indirectly), but also identifiable. As far as the concept of identifiability is concerned, a person is deemed identifiable if he can be identified by any means likely reasonably to be used by the data controller or any other person.
'Processing' is another broad concept defined in the Directive. In fact, every operation performed on data (be it manual or automatic) is *'processing'* in the sense of art. 2(b). As we have seen in the previous sections, MT services perform a series of automatic operations on input data which are far from being a simple word-for-word re-coding. Therefore, it is beyond doubt that MT qualifies as 'processing' of data.
For the purposes of this study, the processing of data in MT services can be divided into two stages. In the first one (that will further be referred to as *'primary processing'*), the user enters data into the service, which are sent to the MT service provider, who then performs a series of operation on the input data and sends the translated output back to the user. At the second stage (that will further be referred to as secondary processing), the MT service provider may process the aggregated input data for different purposes, such as the evaluation and development of the service, statistics or even direct marketing. The following sections will analyse these two stages separately, as they present substantially different legal considerations.

3. Primary processing

For each stage of processing, it is essential to identify the data controller, i.e. *'the person who determines (alone or jointly with others) the purposes and means of the processing of personal data'*. It may seem that as far as primary processing is concerned, the user shall be regarded as the controller, whereas the MT provider is merely a processor (i.e. a person who processes data on behalf of the controller). However, given that the MT provider plays a crucial role in determining the functioning of an MT service, he can also be regarded as a controller. In fact, the definition in the Directive expressly

allows for there being more than one controller for one processing.
The following sections will examine the responsibilities and obligations of both the user and the MT provider.

3.1. Processing by the user

From the user's perspective, two main categories of personal data can be processed at this stage: the data concerning the user himself and the data relating to a third party.
The first case - processing one's own data - does not seem to raise any particular concerns as far as the lawfulness is concerned. In practice, however, the data concerning only the user may be limited to rather narrow circumstances.
Processing of a third person's data may be exempted from the Directive if it is done as a purely personal or household activity. It is not clear how to interpret this category. Textbook examples of such activities include private correspondence and keeping of address books. It may actually seem that even processing for professional, commercial or academic purposes can be covered by the exemption, as long as it is carried out in the course of a purely personal activity (eg. in a private paper notebook, or offline on a personal computer), as the text speaks of personal activities, and not personal purposes. This would suggest that the use of MT tools (in order to obtain an imperfect translation of a text that the user is not personally able to understand, or to translate in the target language), as long as neither the input nor the output data are made public, shall be exempted from the Directive. The scope of the *'household exemption'*, however, has been recently interpreted narrowly by the CJEU in the Rynes case[3]. It is possible, therefore, that the user of an MT service would have to comply with the Directive, especially as far as the grounds for lawfulness of processing are concerned.
The default legal basis for processing should be the data subject's consent. Consent is defined in art. 2(h) of the Directive as *'any freely given specific and informed indication of [the data subject's] wishes by which [he] signifies his agreement to personal data relating to him being processed'*.
This definition does not require that consent be given e.g. in writing[4]. Therefore, for example if the user receives an e-mail in a language that he does not understand, he may imply the sender's consent to enter it into an MT system. In our view, however, implied consent is not easy to apply in this context, as it will likely miss the *'informed'* element, as it is difficult to argue that an average Internet user fully understands the implications of using a 'free' online MT service - and as such cannot validly consent to the processing if this information is not given to him up front. Moreover, even if consent can indeed be implied from the data subject's behaviour, this consent can only concern processing of data relating to the data subject, and not a third person.
It is true that the Directive also allows alternative legal bases for processing, in particular when processing is necessary for the purposes of legitimate interests of the data subject, the data controller or a third party (art. 7(f)). In our view, it is unlikely that the use of MT tools passes the necessity test. In fact, traditional (human) translation (much less problematic from the point of view of privacy) is always possible. The use of MT is therefore never really necessary from the user's point of view.

3.2. Processing by the MT provider

In our view, the only two grounds that can be taken into consideration in the case of data processing by MT providers are: the data subject's consent (art. 7(a)) and performance of a contract to which the data subject is party (art. 7(b)).
As mentioned above, the Directive does allow for implied consent. Such consent can possibly be inferred from the mere fact that the user enters some text in the service and clicks on the 'Translate' button, just like *'dropping a business card in a glass bowl'* can in some limited circumstances be interpreted as consent[5].
Another legal ground that can be thought of in the context of 'free' online MT services is performance of a contract to which the data subject is party. In fact, the MT provider offers an MT service to the user who, by entering data in the service accepts the offer.
The processing of data is therefore necessary for the performance of such a contract -- which in itself may constitute a valid legal basis for processing.
In reality, however, these legal bases are only valid for the processing of data relating to the user. Once again, by processing data relating to a third party, the MT service provider is potentially in breach of the Directive. Just like in the case of processing by the user, processing by the MT provider fits with difficulty within the framework of the Directive.

4. Secondary processing

Some users may imagine that the data entered in a *'free'* online MT service *'disappear'* once the MT process is accomplished. In fact, MT service providers are interested in keeping the data and re-use them in the future.
In fact, the business model behind *'free'* online MT services is simple: they allow to harvest data from users which can then be re-used (either directly by the MT provider or by a third party) for direct or indirect marketing or advertising purposes. Naturally, the data can also be used to improve the tool (by enriching the corpus on which the translation model can be based). In this model, the data (together with additional input from the user) are in fact a form of payment for the service (hence, the services are not really *'free'*).

3 C-212/13, December 11, 2014.
4 See: WP29's opinion 15/2011 on the definition of consent.
5 cf. *idem*

Apart from raising ethical concerns, such behaviour of MT service providers is also doubtful as far as its conformity with the Directive is concerned. Firstly, art. 6.1 (e) prohibits data storage for periods *'longer than necessary for the purposes for which the data were collected'*, which in itself may be a barrier to any form of secondary processing of MT data. Secondly, given that an average user is not even aware of this processing taking place, it is practically impossible for him to exercise rights that are granted to him by the Directive, such as the right of access (art. 12) or, more importantly, the right to object (whose particular instance is the right to be forgotten).

From the point of view of the Directive, two scenarios for *'secondary processing'* (i.e. re-use of data by the MT-providers) should be distinguished: firstly, secondary processing for such purposes as research, evaluation and development of the MT service (translation model); secondly, secondary processing for marketing and advertising purposes. For the sake of simplicity, these two scenarios will be referred to as *'non-commercial'* and *'commercial'* secondary processing.

4.1. Non-commercial secondary processing

In our view, in some cases non-commercial secondary processing may be allowed by the Directive even without additional consent of the data subject. First of all, art. 6.1 (b) interpreted a contrario allows for further processing of data for purposes compatible with the initial purpose, including historical, statistical and research purposes. Therefore, it may seem that the processing for the purposes of statistics and research (including, arguably, the improvement of the translation model) may be allowed. However, according to WP29's opinion one of the key factors in assessing purpose compatibility should be 'the context in which the data have been collected and the reasonable expectations of the data subjects as to their further use'. As explained above, any form of secondary processing of MT data does not seem to meet 'reasonable expectations' of MT users, as most of them simply expect that the data will be deleted after the MT is accomplished.

In fact, MT service providers may be more successful trying to rely on art. 7(f) of the Directive, which allows for processing of personal data *'necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed'*. Indeed, the development of online MT tools is not only in the legitimate interest of the MT service provider, but also in the *'real and present'* interest of the whole community of users. The problem here, however, is that art. 7(f) of the Directive further specifies that fundamental rights and freedoms of the data subject may override other legitimate interests; therefore, the fact that in case of secondary processing users cannot exercise their rights, and in particular their right to be forgotten, may lead a court to reject art. 7(f) as a valid legal ground for such processing.

A special exception for data processing for research purposes is also contained in art. 83 of the General Data Protection Regulation as initially proposed by the Commission; after numerous amendments introduced by the Parliament, its future, however, remains uncertain. If adopted, such an exception would under certain conditions allow for some forms of secondary processing of MT input data.

4.2. Commercial secondary processing

The providers of 'free' online MT data may want to further process the input data for commercial purposes, such as direct and indirect marketing or advertising. It is clear, however, that the Directive does not allow for such form of secondary processing, which enters neither in the scope of art. 6.1 (b), nor art. 7 (f). It would therefore necessitate the data subject's consent, distinct from the one given for primary processing, which this time certainly cannot be implied. In particular, in order to validly consent for such secondary processing, the user would need to be throughly informed. Even if such thorough information is provided to the user, some forms of commercial secondary processing may, in our view, fail to meet the requirement of fairness, distinct from the one of lawfulness (art. 6.1 (a) of the Directive), and therefore violate the principles of the Directive.

5. Conclusions

MT is a very useful and constantly improving technology which may contribute in a very efficient way to crossing the language barrier in digital communications. While the benefits of 'free' online MT cannot be overestimated, the use of this technology is also related to some important privacy risks most users are completely unaware of, and some MT service providers may be tempted to take advantage of this lack of awareness.

The current EU data protection framework, if applied and respected by all the involved actors, does shield the users from most of those privacy risks. However, it some cases it may also place an honest user or an honest provider of these services in danger of breach of law. It should be openly admitted that an online MT service that fully complies with the Directive is possible only theoretically.

6. Bibliographical References

Berger, A. L., et al. (1994). The Candide system for machine translation. In *Proceedings of the workshop on Human Language Technology* (HLT '94). Stroudsburg, PA: Association for Computational Linguistics, pp. 157-162.

Brown P. F., et al. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19, p. 263.

Cribb, M. V. (2000). Machine Translation: The Alternative for the 21st Century? *TESOL Quaterly*, 34(3).

Hutchins, J. W. (1986). *Machine translation: past, present, future*. New York: Halsted Press.

Hutchins, J. W. (1999). Warren Weaver Memorandum: 50th anniversary of machine translation. *MT News International* 22(5).

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Quince, M. (2008). *Why Austria is Ireland*. 1/11/2014. Retrieved from http://itre.cis.upenn.edu/~myl/languagelog/archives/005492.html

Smith, J.R. et al. (2013). Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 1374–1383