Machine Translation for Cross-Language Social Media

Jordi Carrera, Olga Beregovaya, Alex Yanishevsky

PROMT Americas Inc.
330 Townsend St, Suite 204, San Francisco, CA 94107
{jordi.carrera, olga.beregovaya, alexy}@promt.com

Abstract

User-generated content available in weblogs and social media a) contains high noise levels, b) is domain independent, c) is generated fast, d) is available in large quantities and d) is inherently focused on information content and knowledge sharing. Thanks to the new Internet culture, which emphasizes accessibility, openness and active participation, communication needs are less stringent but require faster response and must preserve information content. These properties make user-generated content suitable for machine translation and, more specifically, hybrid machine translation, which combines knowledge representation with statistical modeling. In this article we present a qualitative study of data extracted from the Social Media Dataset: we analyze how naturally occurring phenomena can affect machine translation quality and we show how new hybrid approaches may successfully preserve semantics while at the same time achieving near-optimal levels of linguistic fluency.

Introduction

User generated content

The new Internet and Web 2.0 cultures emphasize both openness and accessibility values, strongly encouraging active participation in knowledge sharing on the part of individual users. It is commonly believed that, by becoming widely generalized, the expression of personal opinions on the Internet will improve our understanding of social dynamics as measured on information mined from social media and social networks (Arguello et al. 2008, Sippel and Brodt 2008, Yi et al. 2003, Pang et al. 2002, Pang and Lee 2004, Wiebe 2000; Yi and Niblack 2005). It is further assumed that this improved understanding will lead to more effective knowledge transfer by allowing to customize content diffusion either in the form of content-based social networking tools or, more specifically, personalized advertising and content-oriented Internet search engines.

The way it naturally occurs in social media and weblogs, user-generated content (UGC henceforth) can be characterized as a) highly noisy, b) domain unrestricted, c) user-centric, d) highly productive (i.e. being generated at a very fast pace in large volumes), and e) inherently focused on information content and knowledge sharing, usually at the expense of correctness in its linguistic codification.

As a result, any approach aiming at reliably extracting information from social media must be designed for 1) large-scale and (as close as possible to) real-time data management, 2) meaning preservation and 3) robustness, which is to be understood not only as tolerance to anomalies in data codification, but also as the system's ability to overcome, when necessary, 4) errors in linguistic formalization and in canonical writing (v.gr. typos, wrong punctuation, unstructured syntax), at least with respect to the codification schemes of traditional top-down content generation approaches found in hierarchical structures with unidirectional rather than bidirectional content flows.

Multilingualism

Although systems successfully designed for features 1), 2), 3) above (i.e. large-scale and real-time data management, robustness and meaning preservation) are expected to be able to deal with UGC to some extent, they still fail to meet the new Internet culture's high standards of content accessibility and openness: right now, the biggest accessibility restriction to digital content remains unaccounted for and consists in linguistic barriers. Even if the "digital divide" as a gap regarding overall infrastructural development is bridged (i.e. a rather "analogical divide", from this standpoint), there still remain as many digital divides for efficient information dissemination as different languages are spoken on the Internet.

As of now, and according to the data of www.internetworldstats.com, the Internet has four-hundred fifty million English speaking users out of one billion five hundred million total users. This means that the market for English language is slightly less than one third of the total market. In other terms, most current approaches to information extraction exploiting social media and UGC (overwhelmingly carried out for and in English) are working with a mere third of the total data available.

All these data are currently being neglected mostly due to the particular characteristics of UGC and its relatively ephemeral character. User-generated content

remains usually untranslated unless the user himself chooses otherwise because 1) it expresses opinions, such that any given user is much more likely to express or to look for a different opinion in the same language than to translate one opinion into or from that language; 2) it is generated and updated at an extremely fast pace and has a very short lifespan, which rules out in practice the possibility of translation by human subjects and 3) is also produced in immense quantities, which, together with the previous point, ends up rendering translation by human subjects effectively impossible (both in terms of time and cost). All this means that there is an enormous body of information being constantly generated which is also being constantly lost behind language barriers: the consolidation of the Web 2.0 has caused an unprecedented increase in the amount of data and each individual user is currently being deprived of most of it.

Yet language diversity in Internet ecology is of paramount importance. As shown in Ramirez-Esparza et al. 2008, an analysis of linguistic data linked to states of depression and extracted from social networking websites revealed that, while English speakers were mainly concerned with medical issues, Spanish speakers were mostly worried about their personal relationships, which provides valuable marketing research and, as it turns out, equally valuable anthropological insights. It has also been shown (Ashforth and Mael 1989, Sippel and Brodt 2008) that social media and, more specifically, blogging, facilitate interpersonal interaction and shape individual identity, while Hewstone et al. 2002 have further demonstrated differential treatment between the *ingroup* and the *outgroup*, referred to as *intergroup bias*. It seems reasonable to generalize their results and expect similar trends to apply to linguistic communities, which would match the empirical observations (Ramirez-Esparza et al. 2008) and would also predict speakers of any given language to be positively biased towards other speakers of the same language. Arguably, this would ultimately anticipate the likelihood for any linguistic community to share specific types of information as a function of its degree of cohesion as measured on the linguistic patterns observed on social media, automatically turning the ability to reach across languages into a competitive advantage for market penetration.

It can thus be realized that linguistic variables are the key to establish content priorities for specific user communities: different languages are a natural result of different social groups with different worldviews and, as a consequence, call for different content. One can take as an illustration the case of the USA Latin American community, whose patterns of spending differ markedly from those of Americans, as reported in studies by Vertis Communication and Brandweek[1,2]. Insofar as a) the spending patterns of people reflect their personal interests and b) their interests influence their linguistic behavior, varying purchasing behavior for any two given linguistic communities will lead to a lack of correspondence in the respective amounts of textual data they generate for their main knowledge domains. Hence it follows some necessary degree of content misalignment in any collection of parallel corpora and, paradoxically, that the main interests of any given community will be those translated the poorest.

Linguistic specificity, on the other hand, also reflects on particular trends of user distribution by language that are in principle unpredictable but that have very specific consequences and give rise to very different translation needs: arguably, no strategic reasons (but just market trends) account for the fact that LiveJournal's user base consists of roughly 65% English-writing and 30% Russian-writing users, yet these numbers (provided by Google Data) make LiveJournal's translation needs very specific with respect to other social networks, and unaccountable on the basis of systems trained on corpora with a different language distribution.

As a consequence, linguistic specificity associated with particular types of content and user communities entails that 1) knowledge extracted from social media analysis for one language cannot be readily extrapolated to other languages or Internet communities, and that 2) tools with an ability to deal with different languages and to do so on a case-by-case basis (i.e. language-by-language) can be expected to yield more relevant results than those lacking such functionality, particularly after the demonstration by (Bautin et al. 2008) that using machine translated content extracted from social media in order to perform sentiment analysis did not harm accuracy and was a largely language independent process.

Therefore, tools exploiting multilingual databases not only have been found to exhibit performance as high as other tools supporting data in one single language, but they have also been shown to be potentially able to access three times as much information and potentially reusable for any number of languages. In fact, UGC analysis software can remain language independent with no reported accuracy loss as long as it is fed the output of a viable machine translation system (Bautin et al. 2008) and even when not using top-of-the-line translation software.

Machine translation

As a result, the output quality of the machine translation (MT henceforth) module in systems performing cross-linguistic data mining becomes a crucial part of any equation aiming at making UGC fully accessible by means of language independent information extraction tools. In this context, two types of machine translation technology are currently available: rule-based machine translation, (RBMT henceforth), and statistical machine translation, (SMT henceforth).

SMT is in fact a direct consequence of the more general trend we have just described consisting in the availability of ever increasing volumes of data, which has now made a number of approaches feasible that were virtually unthinkable of only a few years ago. Unlike

[1] www.marketingcharts.com/television/hispanics-to-buy-more-electronics-than-others-more-influenced-by-tv-internet-5450/vertis-hispanics-research-online-e-june-2008jpg

[2] www.brandweek.com/bw/content_display/news-and-features/packaged-goods/e3i3f22f3dffa4811888f9e647f65157c30

RBMT systems, in which text is processed according to pattern-matching strategies taking advantage of a preexisting knowledge base consisting of linguistic structures and heuristics hand-crafted by teams of language specialists, SMT systems dispose of any such aprioristic knowledge representation and rely solely on quantitative information extracted by systems trained on vast amounts of data. The translations returned by SMT systems are generated by algorithms using some measure in order to establish the degree of similarity between text pairs as recorded on parallel corpora (i.e. databases containing the same linguistic information in two or more languages on a unit-by-unit correspondence, the unit varying depending on the details of any given implementation). As a result, SMT results are usually much more natural and fluent (cfr. examples in Table 3), assuming the system finds a match close enough to the original text. Otherwise, there is no guarantee that the system will return a translation, but only the n-gram bearing the greatest resemblance that can be found in the database.

On the one hand, RBMT as a method can be regarded as materially depleted, due both to having most probably reached its cost-efficiency peak, as well as to the exhaustion of the classic models of knowledge representation (as well as the types of knowledge being represented). Both factors call for more advanced information formalization schemes than are currently available, together with new insights on how to further and more fruitfully exploit available content. This need has triggered the creation of a body of a particularly impressive and pioneering research, but has yielded so far no results able to match the market expectations or to lead to commercially viable implementations.

SMT approaches, on the other hand, have just seen the light of day: text continues to grow by orders of magnitude on a yearly basis and statistical translation engines provide unparalleled adaptability and remarkable robustness to specific types of noise.

Despite its clear advantages, however, SMT has also some obvious drawbacks. First, SMT is much less robust in regards to some other types of noise (v.gr. unstructured syntax and grammatical mistakes) and can in fact lead to important information losses (vid. section *Content loss* below), thus failing to preserve meaning, the main value of all content in general and the key value as far as UGC is concerned.

Second, SMT fails to successfully incorporate another integral part of the user experience of the social media culture, namely, user interaction. Whereas knowledge-based systems provide knowledge representations based on some form of intelligible conceptual categories to which users can contribute their own knowledge and thus achieve both content-oriented customization and, ultimately, more refined information sharing, statistical engines typically lack such a representation and are virtually black boxes with which users can expect no meaningful interaction: SMT systems do return results, but users cannot usefully or significantly influence those results on the basis of their own experience. Instead, they depend on whatever data happen to occur in the corpora the system has been trained on.

In order to compensate for these drawbacks, the latest MT development consists of a hybrid approach combining both the power and coverage of SMT and the knowledge representations characteristic of RBMT. In what follows, we will present the kind of phenomena this new technology is expected to be able to deal with.

Data Analysis

In this section we carry out a qualitative analysis of linguistic data extracted from the Social Media Dataset and present the conclusions following from our study. The methodology applied has been manual inspection, which by its very nature cannot aim at exhaustiveness or representativeness. Instead, our goal is to describe the theoretical motivation of the next generation of hybrid MT engines based on a realistic scenario, and to show the ways in which these can help to achieve the objective of making UGC globally and cross-linguistically available.

We will analyze how naturally occurring phenomena encountered as part of daily user experience with MT software can affect machine translation quality and how new hybrid technology can successfully preserve content while at the same time dealing with noise in order to, ultimately, reach near-optimal levels of linguistic fluency.

In order to be able to establish a comparative analysis and obtain more meaningful results, in the following sections we put side by side the output of two state-of-the-art MT engines, Google Translator for statistical MT, and PROMT's software for rule-based MT.

Phenomena

Consider first the examples in Table 1 (for brevity and ease of exposition, only the English translation has been included here).

No.	Human translation	Promt translation	Google translation
1	The walk in Antwerp begins	There begins (1.1) the walk along (1.2) Antwerp	Start (1.1) the walk in (1.2) Antwerp
2	and what better than to start	and that (2.1) better that (2.2) to start	and what (2.1) better (2.2) to start
3	outdoors enjoying the port of the above mentioned city	outdoors enjoying the port of the above mentioned city	enjoying the outdoors this port city

Table 1. RBMT vs. SMT anomalies on naturalistic data extracted from the Social Media Dataset

In example 1.1, the rule-based translation engine has preferred a rather infrequent syntactic construction over a more common one ("There begins the walk" instead of "The walk begins"). The option returned is not incorrect, however, and is stylistically appropriate, if not likely. The

statistical engine, on the other hand, has taken the verb to be a different grammatical form and has labeled it as an imperative rather than a simple present form.

In example 1.2, on the other hand, the rule-based system incorrectly translated a preposition (*along*), whereas the statistical system did not make the same mistake (*in*). In 2.1, the rule-based engine missed the interrogative "what" due to an orthographic error in the original Spanish text, whereas the statistical engine was able to retrieve the relevant form. In 2.2, however, both the statistical and the rule-based engines failed to identify a comparative structure present in the original (*better than*). The rule-based engine kept the original syntax unaltered enough, though, to allow identifying the immediately following constituent as a separate complement clause (thanks to relying on an a priori knowledge base which is able to impose some degree of structure on the data). As a result, the rule-based engine has been partially able to isolate the relevant concept being compared (*to start outdoors*), whereas the statistical engine has returned a superficially fluent output in which the relevant syntactic position, however, has been inadvertently altered, causing the constituent involved to be more naturally interpreted in a different way (as a verb modifier). This is the same as if a sentence such as "I tried to stay", meaning that what was tried was the fact of staying, was reported to mean that some other unexpressed action was tried in order to achieve the goal of staying. This constitutes a fairly important semantic distinction and failing to account for it may have undesirable consequences on any information extraction system relying on the output of a machine translation module in order to mine relevant content. In this regard, SMT is more likely to cause information losses than RBMT, since only the RBMT proactively exploits pre-existing knowledge representations in order to imbue the data being analyzed with some structure: SMT tends to produce *collage* translations ("*collage* effect" henceforth) under an illusion of linguistic fluency whose meaning may unknowingly be, in reality, further away from the original being input.

Finally, to close the present section, consider example 3: it can be realized from it how a whole sequence of words (*above mentioned*) may eventually disappear in SMT output, whereas the rule-based engine is conservative as regards the source structure and virtually guarantees that all constituents will be translated and none will be deleted or added during the process. Likewise, the word "outdoors" is misplaced by the statistical engine and translated as a modifier of the wrong verb because of its inability to establish the relevant linguistic relationship, namely, what is being enjoyed and how the "outdoors" relates to it. On the other hand, the rule-based engine correctly identified all relevant linguistic relations here, and preserved all significant content in the translation.

Overview

Following from 1.1, it can be concluded that rule-based engines can be successfully customized and account for stylistic variation, whereas statistical engines usually neglect creative language usage in favor of either related but more widespread alternatives, or simply unrelated alternatives for which more data can be found.

Also, according to data such as that in example 1.2, it can be claimed that statistical methods are generally better at coming up with the right prepositions, since prepositional use is normally subject to collocational factors and heavily influenced by lexical selection rather by any abstract part-of-speech categorial patterning such as that used in traditional rule-based approaches.

In addition, following from the data in 2.1, statistical engines show higher robustness to instances of minor spelling errors, typos and general orthographic inadequacy.

As show in 2.2, on the other hand, statistical engines are prone to accidentally affect the meaning of the translation by returning improper output when fed input for which they cannot find sufficiently high-scoring matches (vid. section *Content loss* below). Whereas rule-based rather literally-sounding translations may more often than not seem cluncky, they preserve the constituent structure of the original text and alter the source content the least. Rule-based engines are thus able to frame the overall relational structure of any given translation and can thus generate a linguistic model as close as possible to the original. This model can then serve as output on which a statistical module can then perform smoothing in order to yield more fluent output while, at the same time, preserving all the content.

Finally, the facts in 3 show that, against a widely held belief, word frequency counts are not enough to provide quality translations and, in fact, not even relevant translations. As example 3 demonstrates, and as it can be easily confirmed by performing a Google search (as of February 2009), the bigram "enjoy outdoors" is more than three hundred times as likely as "start outdoors", which, all things being equal, would account for the former's having been preferred as the correct translation. Weighing both these structures merely on the basis of frequency counts, however, is rather inadvisable: "enjoy outdoors" may be much more likely than "start outdoors", but the semantic relations involved are different. In the former bigram, what is enjoyed is the outdoors, while in the latter the outdoors refers to where something else is enjoyed, which yields quantitative comparison of both structures meaningless. This phenomenon can be taken to support the idea that, without any reference to some sort of knowledge representation or structural analysis and on the basis solely of word counts, it is not content that is being translated, but words (i.e. *collage* effect). In the face of this realization, hybrid engines combining statistical modeling with sufficiently rich knowledge representations are again expected to be able to yield increasingly accurate results for the kind of phenomena described in 3, while at the same time retaining all advantages of statistical engines as described in 1.2. and 2.1.

Further exemplification

Noise

As stated in the introduction, UGC may contain varying levels of noise. MT engines performance varies as a function of it. Desirably, MT engines, either RBMT or SMT, should be as robust as possible, though it is usually the case that they are sensitive to noise and each of them either to different types or at different levels. This non-trivially determines their efficiency in order to deal with specific types of data, as shown in the examples in Table 2.

No.	Original	Promt	Google
4	la gran **mayorita** para mal	the big **mayorita** for evil	the vast **majority** for worse
5	es tanto el miedo a seguir en **ste** lugar	it is so much the fear of following in **ste** place	so much fear to stay in **his** place
6	ya no me importa lo **ke** llegue a pasar	the **ke** does not matter for me go so far as to happen	no me importa **ke** rioja comes to pass
7	**pase** un rato bien **chingon** con una linda chica frente a los pandas jajajajajajaja	I **happened** a little bit well **super** with a pretty girl opposite to the pandas jajajajajajaja	**pass** a good time with a nice girl **Chingon** against pandas jajajajajajaja
8	**komienzo** a preguntarme el **porke** en vez del para **ke**	**komienzo** to the **porke** ask me instead of for **ke**	to ask the **komienzo porke** instead of for **ke**
9	**deskonosko** realmente **ke** sucede	**deskonosko** really **ke** it happens	**ke deskonosko** really happens

Table 2. Machine translation examples containing relatively low (4, 5, 6 and 7) and high (8 and 9) amounts of noise

In example 4 and 5, a minor one-character variation (one letter being added to the source text) caused the RBMT to fail to detect a word, whereas the SMT was able to retrieve it. Although implementing a simple pattern-matching algorithm using a similarity measure based on the minimum edit distance allows also RBMT engines to avoid this type of error, it is usually the case that SMT, with its exploitation of large amounts of data where non-noisy instances can easily outweigh noisy ones, makes the whole process considerably easier and to follow in a natural way. It is important to notice, however, that, as shown also in example 5, relying solely on statistical estimates to infer information from noisy data may yield seemingly successful guesses (unsuspiciously fluent, i.e. *collage* effect) that, in reality, are not the right guess (but, at the same time, cannot be told from a right, different one). In this context, it must be considered whether it is more convenient to keep noise in the output for the user to be able to detect it as such and give it a separate treatment, or to smooth it in the first place at the cost of generating some amount of linguistically fluent but misled translations.

On the other hand, in example 6 in Table 2 both the RBMT and the SMT have been unable to identify an item where two out of three characters have been altered. This is unsurprising, since it would amount to sixty percent of the data in any given sample being noise, which can be safely dismissed as rather unusual in any proportionally representative collection of data.

No.	Original	Promt translation	Google translation
10	Se encuentra en la parte norte de la ciudad	It is in the north part of the city	It lies on the north side of town
11	Al inicio del curso	to the beginning of the course	At the beginning of the course
12	entre vosotros tiene que haber alguien más inteligente que yo	between you there has to have someone more intelligent than I	among you there must be someone smarter than me
13	entre las muchas cosas que pasaron	between many things that happened	among the many things that happened
14	Pues bueno,	Since good,	Well,
15	he decidido ponerle fin a un ciclo	I have decided to finish a cycle	I have decided to put an end to a cycle

Table 3. Examples showing the high degree fluency achieved in SMT translation quality with respect to RBMT

This notwithstanding, it is also true that this level of noise may not be too unlikely for particular linguistic items or registers. In these cases, traditional MT tools allow users to deal with recurrent individual instances of noise by providing high levels of customization of the input/output stream, such that noise too statistically marginal to be adequately dealt with by SMT engines can be specifically accounted for by using standard RBMT functionalities. SMT, on the other hand, when encountering noise causing an otherwise frequent structure to become another possible but highly improbable n-gram, may be misled to return the latter's translation on the basis merely of its being the highest available match (which does not mean that it is even a translation). See Table 4 for examples of this phenomenon: in 16 and 17, Spanish neuter pronoun "lo" is found as a part of an n-gram sequence for which no relevant candidate has been found. When faced by these data, the SMT system returns the name of the corresponding province of Spain where the city the acronym "LO" stands for is located. The same applies to examples 18 and 19 as regards Spanish contraction "al": it is translated as the name of a city, "Andalusia", located in the state having "AL" as its acronym, namely, Alabama. In 20, finally, the item "ene" in the source string "ene ste", a misspelling of Spanish "en este" (English "in this"), is translated as "January" because of "ene" being an acronym for the name of the month.

Finally, example 7 in Table 2 exhibits noise levels higher than a RBMT can deal with. Both orthographic and dialectal variation pose here a serious challenge for aprioristic knowledge modeling (the system's expectations are not met, which results in unrecognizable input and accordingly unstructured output). The SMT engine, on the other hand, is able to retrieve from its database a relevant collocation, which in this case successfully accounts for most of the meaning of the original text, at the expense only of negligible grammatical information. Examples 8 and 9, on the other hand, display levels of noise (due to the writer's deliberate intent to break conventional language use for stylistic purposes) currently intractable by MT technology.

No.	Original	Promt translation	Google translation
16	el que mejor se **lo** paso	**the** one that better spent it to himself	**rioja** the best way
17	ya no me importa **lo** ke llegue a pasar	**the** ke does not matter for me go so far as to happen	no me importa ke **rioja** comes to pass
18	averiguie como llegar **al** zoologico de Chapultpec	averiguie like coming to **the** Chapultpec zoo	getting averiguie **andalusia** Zoo Chapultpec
19	El carnicero mata de 10 a 12 perros **al** díÂa	The butcher kills from 10 to 12 dogs to **the** díÂa	The butcher kills 10 to 12 dogs IAWD **andalusia**
20	momentos de ocio que tambien comparto **ene** ste momento	moments of free time that also I share **ene** ste moment	leisure time which also shares his time since **Jan**.

Table 4. Examples showing mistakes made by an SMT translation engine due to statistical fluctuations

Content loss

As already pointed out, empirical impossibility to build a database exhaustive enough to contain all possible n-gram combinations for all individual lexical items in any particular natural language, makes SMT relatively resistant to novel linguistic constructions and to original content. Increasingly large corpora provide increasingly abundant instances of non-original rather than original language use, which may eventually outweigh translations for fragments of new content and dismiss these as being non-canonical language use rather than canonical original use. At best, SMT will remain insensitive to most potentially informative and valuable content (data exhibiting higher levels of entropy in the corpora); at worst, it will naturally tend to omit it, promoting homogeneity rather than knowledge transfer.

Thus, it seems fair to conclude that SMT is most suitable to account for highly frequent items and to achieve high levels of stylistic fluency, but rather unsuitable for preserving content, arguably one of the goals on communication technology in the Information Society, as was already discussed in the introduction.

In this context, and as hybrid technology intends for, rule-based modules can be used to compensate for that drawback: knowledge-rich approaches can proactively apply some specific formalization of previously acquired knowledge in order to generate translations for linguistic units for which no recorded equivalent is available, such

	Original	Promt	Google
21	nada k me haga reaccionar..	anything k make me react.	anything to react to me ..
22	los investigadores encontraron perros vivos en una habitación sin calefacción viviendo el pleno rigor del invierno	the investigators found living dogs in a room without heating living through the full winter rigor	researchers found dogs living in an unheated living room full rigor of winter
23	no me estaba dejando más que decepciones, lágrimas y tristeza	in his west was not leaving me any more than disappointments, tears and sadness	as I was not leaving sunset over disappointments, tears and sadness
24	aquella persona que era tu mundo entero, te ha ido sacando de su vida poco a poco	that person who was your entire world, has been extracting you of his life little by little	he person who was your world, you've been taking your life little by little

Table 5. Examples showing syntax level mistakes made by an SMT translation engine

that statistical modeling will not smooth them nor overrun their content differential as readily as it would accidentally do otherwise.

Table 5 gives examples of linguistic chunks where failure on the part of the SMT engine to find an actual translation matching closely enough the source text has resulted in the closest match's being selected despite its being a false positive.

Thus, in example 21 the RBMT engine successfully translated the subject of the sentence as being the one made to react, whereas the SMT translated something else as reacting to the subject (resulting in a propositional meaning inversion or, put another way, amounting to report a one-hundred percent gain as a one-hundred loss).

In example 22, by misplacing the modifier "living" (originally modifying "dogs", i.e. *living dogs*, but translated by the SMT engine as its syntactic head instead), the SMT crucially failed to capture one of the core ideas in the original (namely, that the dogs were alive against the sentence subject's reported expectations). The RBMT, however, has been able to preserve this meaning component by successfully analyzing the conceptual structure of the source text and being able to map it to its corresponding translation.

Examples 23 and 24, on the other hand, show how the RBMT engine has managed to preserve most of the meaning of the source text. The use of an unrecognized idiomatic expression (i.e. "in its west", which stands for "at its sunset", meaning "at its end") has partially confused the system, but it has then been able again to extract most of the conceptual structure of the rest of the sentences involved, such that something other than the writer has been identified as the grammatical subject. The SMT engine, on the contrary, has translated the writer of the sentence to be also its subject, that is, has translated what a person was experiencing as something being done by that person, which again has caused the loss of a non-trivial part of the content of the source text.

25	SPANISH ORIGINAL	El gobierno chino sostiene que permitirá satisfacer la demanda del Mercado
	HUMAN TRANSLATION	The Chinese government claims that it will allow to meet the Market demand
	PROMT	The Chinese government supports that it will allow to satisfy the demand of the Market
	GOOGLE	The Chinese government claims that will meet market demand
26	SPANISH ORIGINAL	Olga Preciado sostiene que la mayoría de chicos (…) ya ha estado vinculado al activismo politico
	HUMAN TRANSLATION	Olga Preciado claims that most kids have already been linked to political activism
	PROMT	Olga Preciado supports that most of boys (…) have already been linked to the political activism
	GOOGLE	Olga Preciado argues that most kids have already been linked to political activism
27	SPANISH ORIGINAL	Medardo Oleas (…) sostiene que el plebiscito del 28 de septiembre no es una elección
	HUMAN TRANSLATION	Medardo Oleas claims that the plebiscite on September 28 is not an election
	PROMT	Medardo Oleas supports that the plebiscite of September 28 is not an election
	GOOGLE	Medardo Oleas a plebiscite of the claims that September 28 is not a choice
28	SPANISH ORIGINAL	un proyecto de la escuela Santa Rosa de Lima que sostienen las Hermanas de San José de Clunny
	HUMAN TRANSLATION	a project of the school Santa Rosa of Lima supported by the Sisters of San Jose of Clunny
	PROMT	a project of the school Santa Rosa of Lima that there support the Sisters of San Jose of Clunny
	GOOGLE	a project of the Santa Rosa de Lima School argue that the Sisters of St. Joseph of Clunny

Table 6. Examples based on data extracted from the Social Media Dataset

Hybridization

From what has been shown so far, and as suggested at the beginning, the new generation of MT systems is aimed at taking advantage of the strengths and compensating for the weaknesses of both RBMT and SMT approaches in order to address the growing demands posed by both UGC and the blogosphere for more reliable content management tools and, in particular, for MT software able to overcome linguistic barriers.

Hybrid approaches are expected to achieve much higher-quality results by combining the output fluency of statistical modeling with the content preservation ability of the structural analysis carried out by knowledge-based approaches. It is believed that this will help increase MT coverage, quality and robustness, while at the same time not sacrificing accuracy or causing any loss in information content.

As an illustration of the benefits of hybrid MT systems (such as the new generation of PROMT translators) as opposed to both RBMT and SMT separately, consider the examples in Table 6. Table 6 contains all finite forms of the Spanish verb *sostener* occurring in one of the files of the Social Media Dataset, together with the sentences those forms are embedded in. Below each Spanish original, its human translation is given, and the next two lines contain translations by an RBMT and an SMT engine, respectively.

Our concern here is regarding the Spanish verb *sostener* itself, which can be translated into either of two English verbs, *to claim* (possibly also *to argue* in certain contexts) or *to support*. As shown in Table 6, the Social Media Dataset contains three instances of the former sense (25 to 27) and only one of the second (28). As expected, this leads the SMT engine to translate correctly all three occurrences matching the most frequent use of the verb, but to incorrectly generalize this sense as well to the only occurrence of the alternative sense of the verb. The RBMT engine, on the other hand, does not perform substantially better, since it was configured by default to adopt what has turned out to be the least frequent sense. As a result, and for the opposite reasons, the RBMT engine also returns incorrect results. As a matter of fact, it gets most of the occurrences wrong (the most frequent sense) and only one of them right (the only occurrence of the least frequent sense appearing in the corpus).

Thanks to their reliance on a priori knowledge representations and linguistic categories, however, RBMT engines and, as a consequence, any hybrid approach as well, can be customized by declaring a pattern-matching rule that will precondition the translation output on the basis of the conceptual structure of the input.

As can be realized from the data in Table 6, the English translation must be *to claim* whenever the verb is followed by *that*, and *to support* otherwise. Therefore, the linguistic description used by RBMT can be configured a priori and required to meet that expectation, such that whenever the relevant context is encountered, the incorrect translation can be ruled out and the correct one returned instead. Thus, a grammatical rule can be declared whereby *sostener* translates into *to claim* when followed by *that*. This rule will have priority over any empirical data not matching the specified pattern in the system's knowledge base, no matter how statistically weighty: the system will seek to identify the relevant structure by comparing the input with its knowledge base. Upon matching, it will then carry out *smoothing* based on the information available in the corpus, and return the most frequent translation for that specific conceptual structure (instead of merely the most frequent translation of the words involved, i.e. regardless

of their conceptual structure, again consistently with the *collage* effect). As a result, all irrelevant translations will be dismissed if not fulfilling this structural condition, such that the output will not be influenced by mere statistical correlation.

In theory, nothing precludes a pure SMT engine of being able to mimic such an operation. In practice, however, the fact that one of the senses of *sostener* is three times as likely as the other will more often than not outweigh the statistical estimate in favor of the most frequent sense and against the least frequent one, resulting in the former's being translated even when a given occurrence may be referring to the latter sense instead.

Thus, instead of imposing on the translation the linguistic structure of the closest match in the database, as SMT does, hybrid approaches first imbue the source text with linguistic structure and then choose the most likely translation from the available data based on its posterior probability conditioned on the relevant syntactic structure.

Conclusion

In this paper we aimed at showing how linguistic content in social media and in virtually any instance of UGC (weblogs, Internet forums, reviews or social networks) can be made accessible more reliably by means of a new generation of hybrid MT technology. We also aimed at giving reasons why translating UGC is an endeavor in practice unapproachable otherwise: the fast pace at which UGC is produced and the volume in which it is generated cause it to remain untranslated more often than not, which dramatically reduces its value in globalized, multicultural communities.

The goal of this new generation of hybrid MT engines is to create a more language-robust solution that will enhance user experience by allowing something that has been to this point inherently local, such as natural languages, to become global. In this way, large amounts of information may become cross-linguistically available, allowing whole user communities to benefit from accessing intercultural content while at the same time preserving their own identity and thus achieving also cross-cultural integration. All approaches aiming at knowledge extraction from social media rely on the use of words and, more than words themselves, on their content, be it opinion mining, market research, public relations, financial modeling or sentiment analysis. If, as the saying goes, sometimes you cannot find the words, you only have to translate them.

Acknowledgments

The authors wish to thank Anthony Alfonso for his valuable comments and suggestions, as well as for having proofread the draft version of this paper. All remaining mistakes are the sole responsibility of its authors.

References

Arguello, J., Elsas, J.L., Callan, J. and Carbonell, J.G. 2008, *Document Representation and Query Expansion Models for Blog Recommendation*. Proceedings of the International AAAI Conference on Weblogs and Social Media.

Ashforth, B. E. and Mael, F. 1989. Social identity theory and the organization. *Academy of Management Review.* 14(1): 20-39.

Bautin, M., Vijayarenu, L. and Skiena, S. 2008. *International Sentiment Analysis for News and Blogs.* Proceedings of the International AAAI Conference on Weblogs and Social Media.

Hewstone, M., Rubin, M., and Willis, H. 2002. Intergroup bias. *Annual Review of Psychology.* 53: 575-604.

Pang, B. and Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification usingmachine learning techniques. In *EMNLP'02*.

Pang, B. and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 271–278.

Ramirez-Esparza, N., Chung, C.K., Kacewiz, E. and Pennebaker, W. 2008. *The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches.* Proceedings of the International AAAI Conference on Weblogs and Social Media.

Sippel, B. and Brodt, S.E., 2008. *The Psychology of Blogging Communities: Social Identities and Knowledge Transfer Across Work-Groups*. Proceedings of the International AAAI Conference on Weblogs and Social Media.

Wiebe, J. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*, 735–740.

Yi, J. and Nasukawa, T.; Bunescu, R.; and Niblack, W. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *ICDM '03*, 427. Washington, DC, USA: IEEE Computer Society.

Yi, J., and Niblack, W. 2005. Sentiment mining in webfountain. In *ICDE'05*, 1073–1083.Washington, DC, USA: IEEE Computer Society.