Connotation in Translation

Marine Carpuat
Department of Computer Science
University of Maryland
College Park, MD, USA
marine@cs.umd.edu

Abstract

We present a pilot study analyzing the connotative language found in a bilingual corpus of French and English headlines. We find that (1) manual annotation of connotation at the word-level is more reliable than using segment-level judgments, (2) connotation polarity is often, but not always, preserved in reference translations produced by humans, (3) machine translated text does not preserve the connotative language identified by an English connotation lexicon. These lessons will helps us build new resources to learn better models of connotation and translation.

1 Introduction

Subtle shades of meaning beyond surface meaning are receiving increasing attention in Natural Language Processing. Recognizing that even words that are objective on the surface can reveal sentiment of the writer or evoke emotions in readers, Feng et al. (2013) show that the connotation of words can be induced from corpora in an unsupervised fashion, and that the learned connotation polarity of words is useful for sentiment analysis tasks. While such connotation resources only exist for English at the moment, sentiment and subjectivity analysis (Pang and Lee, 2008) increasingly addresses other languages (Banea et al., 2011).

This leads us to ask whether connotation can also be studied in the cross-lingual and multilingual setting. Modeling and detecting differences of connotation across languages would have many applications, e.g., enabling comparison of social media discussions in different languages. But since connotation is a more subtle form of meaning, with cultural and emotional associations, it is not clear to what extend we can expect it to be preserved in translation. On the one hand, we expect correct translations to preserve the meaning of the source: this is the key assumption underlying alignment algorithms in statistical machine translation (Brown et al., 1990), as well as the use of translations to capture the meaning of words in lexical semantics (Resnik and Yarowsky, 1999; Callison-Burch, 2007; Apidianaki, 2009; Carpuat, 2013, among others). On the other hand, cross-lingual structural divergences (Dorr, 1994) might introduce subtle but unavoidable shifts in meaning as part of the translation process.

In this short paper, we report on a pilot study on connotation and translation, using human and machine translated text, and manual as well as automatic tagging of connotative language. We will see that connotation is often, but not always, preserved in translation. This suggests that new models will be needed to represent, predict and use word connotation in more than one language.

2 Defining connotation

We adopt the notion of word connotation defined, and used, by Feng et al. (2013). Connotation refers to "an idea or feeling that a word invokes in addition to its literal or primary meaning [or denotation]." Words with positive connotation describe "physical objects or abstract concepts that people generally value, cherish or care about", while words with negative connotation "describe physical objects or abstract concepts that people generally disvalue or avoid".

As a result, connotation can be evoked by words that do not express sentiment (either explicitly or implicitly), and that would be considered neutral in a sentiment analysis or opinion mining task. For instance, the nouns "life" and "home" are annotated as objective in SentiWordNet (Baccianella et al., 2010), while they carry a positive connotation according to the definition above.

3 Study conditions

Languages: We choose French and English as tar-

Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 9–15,
Lisboa, Portugal, 17 September, 2015. ©2015 Association for Computational Linguistics.

get languages, as these are resource-rich languages and machine translation between them can be achieved with reasonably high quality (Callison-Burch et al., 2009; Bojar et al., 2013).

Domain: We collect text from the Global Voices[1] website Unlike more traditional news sources, Global Voices content is produced by a community of volunteers who curate, verify and translate trending news emerging from social media or blogs. We crawled Global Voices to collect articles that are translations of each other. This study focuses on headlines from these articles: we anticipate that headlines are good candidates for studying connotative language since they aim to provide a concise summary of a news story, and are often written to capture the attention of readers.

Size: We work with a sample of 245 parallel headlines, and study the connotation in each language using both automatic and manual analysis.

4 Does machine translation preserve connotative language?

We start our analysis of connotation using fully automatic means: machine translation and an automatically induced connotation lexicon. We use the lexicon to tag connotative words in both human-produced English, and machine-translated English. If machine translation preserves connotation, we expect to find a high overlap between connotative words in machine translated text and the human-produced reference, and we expect the connotation polarity to remain the same.

4.1 Marking connotative language

We use the English connotation lexicon[2] to tag connotative language. We run the Stanford part-of-speech tagger on all our English examples (Toutanova et al., 2003), and tag word and part-of-speech pairs that are found in the lexicon with their polarity (i.e. negative, positive or neutral). [3]

For instance, in the example "Guinea-Bissau: Citizen Frustration and Defiance in Face of Turmoil", the connotation lexicon detects one word with positive connotation ("citizen_NN") and three words with negative connotation ("frustration_NN", "defiance_NN" and "turmoil_NN").

This broad-coverage lexicon was automatically induced from raw text, based on the intuition that connotation can be propagated to the entire vocabulary based on co-occurrences with a small set of seed connotative predicates of known polarity. For instance, the arguments of "enjoy" are typically positive, while those of "suffer" are typically negative. Follow-up work showed that connotation can be associated with fine-grained word senses(Kang et al., 2014), but we limit our analysis of connotation at the word level at this stage.

4.2 Machine translation systems

We produce automatic translations of the French headlines into English using two different machine translation systems.

First, we use Google Translate, since this free online system is known to achieve good translation quality in a variety of domains for French to English translation. Second, we build a system using publicly available resources, to complement the black-box Google Translate system. We use the hierarchical phrase-based machine translation model (Chiang, 2005) from the open-source cdec toolkit (Dyer et al., 2010), and datasets from the Workshop on Machine Translation.[4]

Google Translate achieves an uncased BLEU score (Papineni et al., 2002) of 20.13, and the cdec-based system 14.60. The lower score of the cdec system reflects the nature of its training data which is primarily drawn from parliament proceedings rather than news, as well as the difficulty of translating headlines. The translation quality is nevertheless reasonable, as illustrated by the randomly selected examples in Table 1.

4.3 Connotative words in human vs. machine-translated text

First, we note that connotative language is found in 89% of the original English examples and 92% of the machine-translated examples. This confirms our intuition that Global Voices headlines are a good source of connotative language.

Second, we compare the connotative language found in machine translated text to the connotative language found in the reference translations

[1] `https://globalvoicesonline.org/about/`

[2] `http://www3.cs.stonybrook.edu/~ychoi/connotation/data/connotation_lexicon_a.0.1.csv`

[3] Words that are out of the vocabulary of the connotation lexicon are considered neutral in this experiment.

[4] Our training set comprises more than two million segment pairs from Europarl and News Commentary data from `www.statmt.org/wmt15`, and our English language model is trained on the additional English news corpora. Translation hypotheses are scored using standard features, including a 4-gram language model. We tune using the MIRA algorithm.

	human references vs. machine translation
input	Visages de la crise et appels au secours
reference	Faces of the crisis and a cry for help
google	Faces of the crisis and calls for help
euro+news	faces of the crisis and calls to the rescue
input	Record de financement collectif pour un documentaire sur lindpendance de la Catalogne
reference	Crowdfunders Empty Pockets for Catalan Independence
google	Collective fundraising record for a documentary on the independence of Catalonia
euro+news	collective record funding for a documentary on catalonia s independence

Table 1: Machine translation output for two systems: (1) Google Translate (Google), and (2) a hierarchical phrase-based system trained on WMT data (euro+news).

Translation System	Google	euro+news
Do positive words overlap with references?		
Precision	42.35	56.13
Recall	30.03	53.87
Do negative words overlap with references?		
Precision	50.75	52.60
Recall	46.69	50.53
Do content words overlap with references?		
Precision	37.35	49.70
Recall	41.56	58.52

Table 2: Are connotative words in machine translation output found in reference translations?

produced by humans. We use Precision and Recall metrics to represent the overlap.

Table 2 shows that precision and recall are in the 40-50 range for connotative words, indicating that they are often not found in the reference. However, this happens less frequently for connotative words than for content words in general: precision with respect to the reference words is higher for connotative words (positive or negative) than for all content words.

Surprisingly, the translations obtained using our in-house system achieves a higher overlap with references despite having a lower translation quality according to BLEU. This might be explained by the nature of its training data which is presumably smaller and more homogeneous, resulting in translations that might be more literal at the cost of fluency, resulting in more matches for content words, and fewer matches for the higher order n-grams taken into account in BLEU.

4.4 Segment-level connotation analysis

Lastly, we use the polarity of the words to compute the dominant polarity of the entire headline. Following the sentiment analysis experiments of (Feng et al., 2013), the dominant polarity is defined as $pol(E) = argmax_{pol=pos,neg} \sum_{e \in E_{pol}} w_{E_{pol}}(e)$ where $w_{E_{pol}}(e) = 2$ if $e \in E_{pol}$ and e is a verb or an adjective; $w_{E_{pol}}(e) = 1$ if $e \in A$ and e has another part-of-speech. Based on this statistic, the dominant connotation of the reference English vs. machine translated English headlines only agree in 60% of the examples considered.

Translation System	Google	euro+news
Comparing MT to positive references		
Precision	54.34	54.34
Recall	30.86	28.73
Comparing MT to negative references		
Precision	62.40	50.40
Recall	75.72	76.82

Table 3: Comparing the dominant connotation of the entire machine translated segment to that of the reference for our two systems.

Taken together, these results suggest that machine translation does not preserve connotative language accurately, even for an “easy” language pair such as as French-English. This differs from prior work on sentiment analysis which suggests that even imperfect machine translation can be good enough to port systems from e.g., English to Arabic dialects (Salameh et al., 2015), or to project labels of subjectivity from English into Romanian and Spanish (Banea et al., 2008).

However, our study of connotation differs from prior work in two important ways: (1) as defined in Section 2, *connotation* refers to meaning that is evoked or associated with a word, while *sentiment or subjectivity* tends to be more explicit. So we expect connotation shifts to be more subtle. (2) our study focuses on *word* connotation, while prior cross-lingual analyses have focused on sentiment/subjectivity at the *segment* level, and are therefore expected to be more tolerant of machine translation errors.

5 Human connotation judgments on human-translated examples

We now turn to manual evaluation of connotation expressed in French and English using manually translated data.

5.1 Defining an annotation scheme

We collect human judgments for the connotation of a given headline. Each annotator is asked whether the language used in the headline implies (1) something positive, (b) something negative, (c) both, or (d) neither (neutral), according to the definition of connotation and its polarity from Section 2. Annotations were produced by native speakers independently for each language, using two different schemes and sets of instructions.

Segment-level 3-way annotation At first, annotators were asked to mark whether the dominant connotation of each segment (i.e. the complete headline) is positive, negative, or neutral. This task was inspired by prior segment-level annotation schemes used for annotating more overt emotion and its polarity in news headlines (Strapparava and Mihalcea, 2007). The inter-annotator agreement (Cohen, 1960) was poor between the two versions of the English annotations, and even worse between annotations of French and English text (see Table 4).

Kappa	en 3a	en 3b	fr 3a
en 3a	100	67.20	55.20
en 3b	67.20	100	55.31

Table 4: Inter-annotator agreement for segment-level 3-way annotation of connotation (positive vs. negative vs. neutral)

Bag-of-words 4-way annotation We then redefined the annotation scheme to discriminate between language that is neutral and language that contains both positive and negative connotations. This yields a set of four labels. We call this annotation “bag-of-words” because it simply indicates whether there exists words in the segment with negative or positive connotation, instead of attempting to assign a single dominant connotation label to the entire segment. This schemes results in higher agreement as measured by Kappa score (Cohen, 1960), both within and across languages (see Table 5).

Kappa	en 4a	en 4b	fr 4a	fr 4b
en 4a	100	73.79	71.08	70.35
en 4b	73.79	100	73.54	72.28
fr 4a	70.35	72.28	100	80.07

Table 5: Inter-annotator agreement for bag-of-word 4-way annotation of connotation (positive vs. negative vs. both vs. neutral)

The “both” category allows annotators to avoid difficult decisions for the confusing examples where positive and negative words are observed in the same examples (see Table 7). The agreement within languages remains higher across languages.

Label	Example
neu	Russia: Online Cooperation as an Alternative for Government?
pos	Russie : la collaboration en ligne comme nouvelle forme de gouvernance ?
neg	China: Wiping Sweat Gate
neu	Chine : le commissaire politique essuie la sueur du front des policiers
pos	China: The Most Awesome Train Door
neu	Chine : Métro de Pékin, attention à la fermeture des portes !
neu	Nicaragua: Opposition Youth Affected by Hacktivism
neg	Nicaragua : Les jeunes de lopposition victimes de piratage informatique

Table 6: Agreement within and Disagreement across languages: negative (neg); positive (pos); both (both); neutral (neu)

5.2 Agreement within and across languages

While we expect the annotation task to be difficult, we found that agreements are more frequent than disagreements both within and across languages.

In fact, all four annotations are identical for 71% of examples, which suggests that the majority of the headlines are not ambiguous. Such examples of agreement can be found in Table 7. English annotations disagree for 16.8% of examples; while French annotations disagree only for 12.30%.

5.3 Disagreement within and across languages

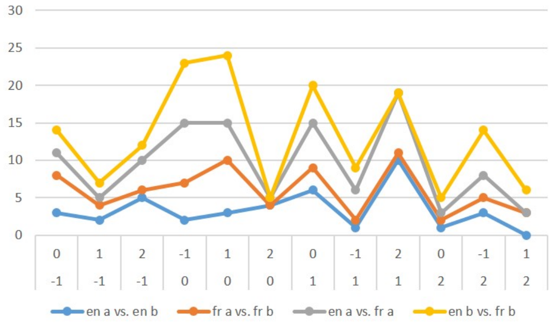

Figure 1: Disagreement in pairwise comparison of annotations: the x axis represents disagreement for each label pair (-1 = negative; 1 = positive; 2 = both; 0 = neutral), the y axis represents the number of observed examples.

Figure 1 summarizes the disagreements observed within English and French annotation, as

Label	Example
neg	Uganda: Government Quiet as Famine Takes Toll
neg	Ouganda : Le gouvernement garde le silence sur la famine
pos	Mexico: Indigenous Long-Distance Runner Wins International Race
pos	Mexique : Une femme de la tribu Tarahumara remporte une course internationale
both	Spain: 12M, a Ray of Sun in the Midst of the Crisis
both	Espagne : Le premier anniversaire des Indignés, un rayon de soleil en pleine crise
neu	China: Graduate thesis or practical training?
neu	Chine : Vaut-il mieux avoir une thèse ou une formation pratique ?

Table 7: Agreement within and across languages.

well as across languages. We observe that there are fewer disagreements between monolingual annotations than across languages. The most frequent confusion is between "positive" or "both" in monolingual, while confusions between "neutral" and "positive" as well as "neutral" and "negative" increase in cross-lingual comparisons.

For a small number of examples (4.5%), French and English annotations are internally consistent within each language but disagree across languages. This happens when one example is deemed neutral or considered to have both negative and positive polarity in one language, but is considered only positive or negative in the other. A sample of such examples is given in Table 6. The differences are due to a number of factors. In the most extreme case, we have an idiomatic expressions with a strong connotation polarity, such as the English suffix "Gate" used to denote a political scandal (derived from "Watergate"). This suffix does not have a direct equivalent in French, and the translation loses the strong negative connotation present in the English. More frequently, key words that convey connotation are translated with words that have a weaker connotation (e.g. the strongly negative "victimes" becomes the more neutral "affected", the positive sounding "serendipity" is dropped from the French version of the headline.)

6 Automatic predictions vs. human labels

Finally, we compare the automatic predictions based on the connotation lexicon from Section 4 to the manual annotation of connotation collected in Section 5. To focus on the most reliable annotations, we only use the subset of examples where intra-language annotations are consistent, which yields a smaller subset of 232 examples out of the initial 244. Furthermore, for each example, we compare the positive vs. negative connotation strengths from Section 4, so as to predict one of the four classes for each example.

reference	input	accuracy
fr	en	44.39
en	en	46.12
fr	mt	40.94
en	mt	37.93

Table 8: Connotation lexicon predictions on English headlines

A baseline predicting the most frequent class ("negative") would get an accuracy of 55%. So the main lesson of this comparison is that using the lexicon out of the box is not sufficient to replicate the decisions of human annotators. Nevertheless, it is reassuring that predictions based on the English headlines agree more with English annotations, while predictions based on machine translation of French agree more with manual annotations of the original French.

7 Discussion

We have studied the connotation of French and English headlines using both manual and automatic annotations and translations.

The manual annotation revealed that translations can diverge in connotation, even in manually translated parallel texts in closely related languages. This suggests that further cross-lingual studies should not use parallel corpora to project annotations blindly. Perhaps more importantly, we found that annotating connotation reliably requires working with a set of four categories ("positive", "negative", "both" or "neutral") to achieve better inter-annotator agreement. We will use these lessons to collect and annotate larger datasets with more annotators, and more languages.

As can be expected, simple lexicon-based predictors are far from sufficient to determine the dominant connotation of a segment. This is consistent with the observations of (Greene and Resnik, 2009) who developed syntactically motivated features for the analysis of implicit sentiment. Accordingly, we will focus on developing better models of connotation preservation and divergence across languages in the future.

References

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 77–85, Athens, Greece, March.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC)*.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 127–135, Stroudsburg, PA, USA.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2011. Multilingual sentiment and subjectivity. *Multilingual Natural Language Processing*.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August.

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederik Jelinek, John Lafferty, Robert Mercer, and Paul Rossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh.

Marine Carpuat. 2013. A semantic evaluation of machine translation lexical choice. In *Proceedings of the 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Atlanta, USA, May.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270, Ann Arbor, Michigan.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1):37–46.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784, Sofia, Bulgaria, August.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 503–511, Stroudsburg, PA, USA.

Jun Seok Kang, Song Feng, Leman Akoglu, and Yejin Choi. 2014. ConnotationWordNet: Learning connotation over the Word+Sense network. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1544–1554, Baltimore, Maryland, June.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July.

Philip Resnik and David Yarowsky. 1999. Distinguising systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado, May–June.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 70–74.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North*

American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, pages 173–180, Stroudsburg, PA, USA.