# Research on the Building of Emotion Metaphor Corpus Based On Machine Translation

Xiaona Jiang
College English Teaching Department, Henan University
Kaifeng, Henan, 475001, China
jxn@henu.edu.cn

*Journal of Digital Information Management*

***ABSTRACT:*** *Metaphor is a result of emotion conceptualization hidden in human language. Due to the complexity and abstraction of human emotion, it is difficult to calculate and create a model for it. However, emotion modeling and calculation is of great significance in the process of machine translation (MT). Usually, incorporating the calculation of emotion metaphors in machine translation could make the language much more vivid and meet the standards of faithfulness, expressiveness and elegance in translation. Normally, calculation of emotion metaphor adopts machine learning and pattern identification, and it requires the samples from emotion metaphor corpus of large scale and high quality. The thesis builds an English and Chinese bilingual corpus with affluent emotion metaphors and supports data to emotion metaphor calculation by machine translation. In the process of building emotion metaphor corpus, 5 main procedures including theoretical framework, design principles, data collection, data annotation and index monitoring are illustrated. Finally, machine translation experiment has been done in emotion metaphor corpus built in this thesis, which adopts same recurrent neural network and LSTM mnemon to compare with existing machine translation corpora. Result shows that the emotion metaphor corpus built in this thesis is able to express emotion metaphor in machine translation.*

**Subject Categories and Descriptors**
**[H.1.2 User/Machine Systems] [I.2.7 Natural Language Processing];** Machine translation

**General Terms:** Machine Learning., Bilingual Corpus, Translation, Pattern Identification, Emotion metaphors

**Key words:** Emotion Metaphor, Machine Translation, Corpus, Recurrent Neural Network, LSTM Mnemon

**Received:** 15 March 2017, Revised 24 April 2017, Accepted 2 May 2017

**DOI:** 10.6025/jdim/2017/15/4/224-229

## 1. Introduction

With the burgeoning development of computer and artificial intelligence technology, human has stepped into an era of information explosion. The languages' translation, which is the oldest communication method for people, has ushered in an era of machine automatic translation [1]. At present, large amount of languages contain various kinds of emotional descriptions and judgments, and the feelings can only be understood by deep understanding and processing. In the process of machine translation, people are eager to empower the machine with capabilities of feeling, understanding and transferring emotion in translation, and adopting smart algorithm to calculate and imitate emotions in massive data automatically [2]. Emotion calculation is one of the important researches in natural language processing (NLP). Calculating emotions of different samples in the corpus is key to stepping into deep semantic gap of computer linguistics and machine translation, which has important research meaning.

In natural languages, no matter Chinese or English, people always express their emotions by colorful metaphors. These corpora are abstract, unclear and hard to express by formalized symbols, such as, "兴高采烈" (xìng gâo cǐi liè, elation), "战战兢兢" zhàn zhàn jîng jîng, jitter), (jié rán yì shçn, alone) "孑然一身", 悠闲自在" (yôu xián zì zài, leisurely and carefree), etc. Emotion metaphor means emotional expressions hidden in the natural languages, which takes up over 1/3 of our conversations in our lifetime. Everyone uses more than 21 million times of emotion metaphor to make expression in his or her whole life. Based on the physiological experience [3], emotion metaphor is the conceptualized expression of complex and changeable emotions, and a major representation. Calculation for emotion metaphor is important in machine translation. At present, emotion metaphor has hardly been covered in machine translation corpus. The thesis gives data support to machine translation research of emotion metaphor by building emotion metaphor database of high quality and large scale.

## 2. Current Situation of Machine Translation's Emotion Metaphor Corpus

There are few emotion metaphor corpora for machine translation now. The first emotion metaphor corpus is Master Metaphor List (MML). This corpus was built under the guidance of metaphor theory, hence subsequent emotion metaphor corpus can find their trace in MML. MML corpus drew metaphor examples from books and documents, online BBS and students' articles. Emotion metaphors in the building of this corpus were divided into 4 categories: time, emotion, mentality and the others. MML corpus has built an effective mapping classification and theoretical basis by mapping examples from source zone to target zone by these 4 categories, but theoretical framework of metaphor's lexical structure is still unclear.

Unclear definition of lexical structure of MML will lead to insufficiency of words in machine translation and cause series of problems. To improve the MML, Metalude corpus collects more than 9,000 emotion metaphor examples on the basis of researching emotion metaphor by words, these examples include information such as noumenons, vehicles, literal meanings, word class and classification, etc. The improvement could solve the biggest problem of MML corpus, but Metalude did not divide emotion metaphor in detail, which led to abstract research direction. Besides, Metabank was based on collecting, generalizing and expanding traditional corpora, it collected newspaper and mail text in MML and expanded their metaphor contents but did not pay attention to emotional factors of metaphor. Therefore, this corpus only fits metaphor research in specialized fields. The emotion metaphor corpus is the only topic of this thesis. In terms of the building of emotion metaphor corpus in Chinese, some major corpora include: Chinese sentence-based metaphor corpus built by Xiamen University which contains more than 10,000 Chinese metaphor sentences that could be calculated but does not support emotion metaphor calculation just like Metabank, the tourist information emotion metaphor corpus by Tsinghua University, and information retrieval emotion metaphor corpus by Dalian University of Technology, etc. However, these Chinese corpora have limited corpus amount and emotional calculation. To apply emotion metaphor corpus in E-C and C-E machine translation more efficiently, it is pressing to build a detailed emotion metaphor corpus.

The building of emotion metaphor corpora at present shows that we can take pages from foreign corpora building since they are comprehensive and thorough, but their emotion metaphor calculation and metaphor classification system construction are uneven. Domestically, Chinese emotion metaphor corpora target limited audience, and draw from different sources, thus lack coordination and unified standard. To solve the present problems of corpora of C-E and E-C machine translation, the thesis aims to build a detailed emotion metaphor corpus with such characteristics as:

(1) Fit for emotion metaphor calculation. Emotion metaphor can be recognized by building large scale and detailed emotion metaphor corpus, non-emotion metaphor corpus, field classifying, and predicting the corpus based on distances of different semantics.

(2) Build machine translation connection for both E-C and C-E translation. Generally, Chinese and English differ in intension and metaphors. To promote the mutual machine translation between English and Chinese, the thesis tags metaphor grounds of machine translation between English and Chinese when building the corpus. Connecting translation between English and Chinese by grounds, and researching connections of tagged grounds are conducive to exploring metaphor differences between English and Chinese and generating better machine translation results.

## 3. The Building of Emotion Metaphor Corpus

### 3.1 Theoretical Framework

The theoretical frameworks for building emotion metaphor corpus require defining emotional boundary, emotional categories, noumenons and vehicles according to emotional calculation theory and principles. According to the present researches, emotional classification and boundary are unclear, which have been divided into 8 or 6 categories and Chinese experts divided that into 7 broad headings and 20 subdivisions. To build a better theoretical framework of emotional boundary and categories, the thesis divides emotions into 7 broad headings: happy, fine, anger, sad, fear, hate and surprise, and 22 subdivisions based on sentiment classification method in References [9]. In the emotional theoretical framework, emotions are divided into broad headings and subdivisions to facilitate the modeling of emotion metaphor computing.

Apart from emotional boundary and categories, the building semantic domain is also of great importance. Semantic domain can be seen as a process when known concepts

are mapped into unknown concepts, and such process is called category dislocation. Tagging semantic domain is key to making emotion metaphor calculation and offering monitorable tags to emotion metaphors' identification, which is conducive to building monitorable calculation models. The thesis refers the classifications [10] from "Roget's Inter-national Thesaurus" during the division of semantic domain which includes 5 levels and 27 precise semantic domains.

### 3.2 Design Principles

The emotion metaphor corpus in the thesis aims to serve emotion metaphor calculation of machine translation, and it consists of two parts including offline work and online work. The offline work requires artificial tags for existing corpus data, while online work requires collecting data during the use of corpus by machine learning algorithm and making tags automatically by machine learning classification. The design principles include the following 5 parts [11]:

(1) The collection of emotion metaphor data should adopt unified strategies and principles.

(2) The process of emotions tagging should adopt unified strategies and principles.

(3) Quality supervision and feedback system are required to reflect present corpus's contents, tags and quality.

(4) The design should have a searching system where the whole mapping process could be searched by using present noumenon, metaphor or ground as key words.

(5) A friendly interactive mode is required to offer useful data support for machine translation's mathematical modeling.

### 3.3 Data Collection

Data collection is a basic part of building a corpus. Data's types and contents need to ensure these two aspects during the data collection: first, diachrony or synchrony; second, plentiful emotional information[12]. Diachrony and synchrony are basic characteristics of a corpus, data with these characteristics could dig out the forming reasons, locus and change regularities of emotion metaphors. Some emotional words, such as "happy'" or "sad", need to be understood with their diachrony characters like historical reasons and locus to understand emotion metaphors inside them. Some other words, such as "afflicted with all ills " and "bursting with happiness", can only be understood when their emotional mapping's synchrony is provided. Therefore, an emotion metaphor corpora only becomes qualified when its data for building emotion metaphors include diachrony or synchrony.

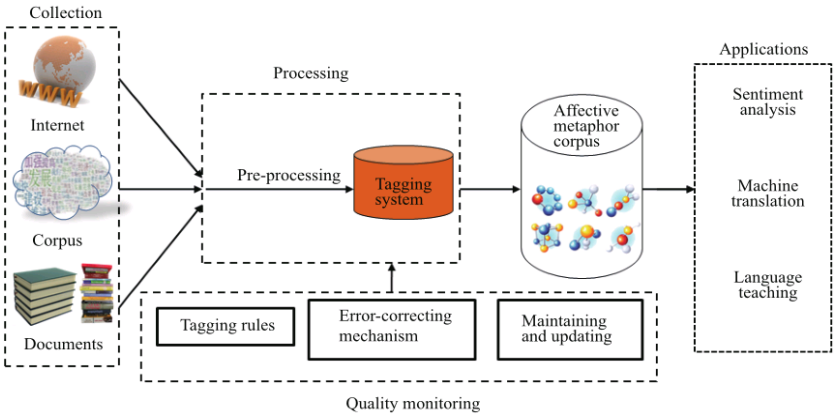



Figure 1. Shows the whole process of design principles of the emotion metaphor corpus

| Sources | Details | Words | Sentences | Pages |
|---|---|---|---|---|
| Books(Metaphors) | " Metaphors We Live by", About 11 books | 18378 | 3828 | 282 |
| Dictionaries | " A Dictionary English Metaphors", 3 traditional dict. | 22193 | - | - |
| Books(fairy tales) | "Grimms' Fairy Tales", about 3 books | 77362 | 4982 | 182 |
| Movie scripts | "Life is beautiful", about 12 movie scripts | 124837 | 10393 | 349 |
| Journals and magazines | "Youth Literary Digest", about 36 Journals and magazines | 41828372 | 229383 | 3494 |
| Textbooks | "College English ", about 22 textbooks | 382712 | 13902 | 489 |
| Micro-blog | Whole comments of micro-blog | 1892838 | 104948 | 39392 |
| Corpora | About 30 corpora | 1584787 | 139028 | 38941 |

Table 1. Source Information of the Emotion Metaphor Corpus

Aside from diachrony and synchrony, source of corpus must be ensured as well. Corpora in this thesis are limited from the Internet, books and documents and present existing corpora which include textbook, literary works, online comments and poets, etc in recent three years. Table 1 provides source information of emotion metaphor corpus.

### 3.4 Data Annotation

The building of corpus is used to assist machine to learn mathematical modeling, then let machine translation achieve emotion metaphor by mathematical models. Mathematical modeling needs learning methods with supervision, which means all emotion metaphors of corpora need relevant annotation or tag. As for the offline artificial tags, the thesis adopts TEI (text encoding initiative) [13], and selects useful tags number and their descriptions of corpora to tag the collected information. TEI is one of the efficient tagging methods to solve tagging ambiguity, it is simple, efficient and helps to improve emotion metaphor's tagging accuracy, consistency and efficiency. The basic framework of TEI is:

$$MetaphorModel = (tenor, vehicle, ground, [indicator], category, emtion\ [note]) \quad (1)$$

In this formula, tenor is the subject to which attributes are ascribed, vehicle is the subject from which the attributes are derived, ground is the source from which we draw metaphorical expressions, indicator is what intrigues the emotion, category is the classification of emotions, emotion is the emotion boundary, and note is the short comment of a word.

For the emotion metaphor tags of Chinese and English, the thesis gives the following examples.

Chinese example:"孙晓红在大门口笑颜如花，像波光粼粼的湖面.
(Sûn XiÎohóng zài dà mén kÒu xiào yán rú huâ, xiàng bô guâng lín lín de hú miàn, Xiaohong Sun smiles at the gate, and her smile is like a glittering lake.)

[Xiaohong Sun smiles, B, 11111 ]
[a glittering lake, Y, 12312 ]
[happy, D ]
[PA]
[is like, D]
English example: Jimmy is as merry as a ice ball.
[Jimmy, B, 11111]
[ice ball, Y, 12315]
[merry, D]
[PA]

### 3.5 Index Monitoring

Index monitoring guarantees the quality of corpus. 7 volunteers accomplished the tagging task of this thesis, and the tagging team included one English teacher, two English post-graduates, one Chinese teacher, two Chinese post-graduates and one computer-major post-graduate. During the tagging, post-graduates were divided into teams and teachers conducted cross check. When cross check had no differences, the result was tagged as correct. When cross check had differences, the differences were recorded and discussed by these 7 people to ensure the correctness and consistency. During the tagging, 87% data had no differences, 13% data had difference and were discussed later. In the end, words were appropriately denoted.

Meanwhile, an error-correcting mechanism was adopted in the process of online automatic tagging. Such mechanism refused any input when tagged metaphor emotions were not in accordance with emotions of tagged corpus's sentences and word noumenon. The error-correcting mechanism is shown as follows:

$$flag = WordConsistency(M_{emo}, W_{emo}) \cap SentConsistency(M_{emo}, W_{emo}) \quad (2)$$

When tagged results are in accordance with source words' emotions, $WordConsistency(M_{emo}, W_{emo})$ and $SentConsistency(M_{emo}, W_{emo})$ are registered as 1, otherwise, as 0. The tagged results will pass the error-correcting mechanism and be recorded into the corpus only when "flag" is 1.

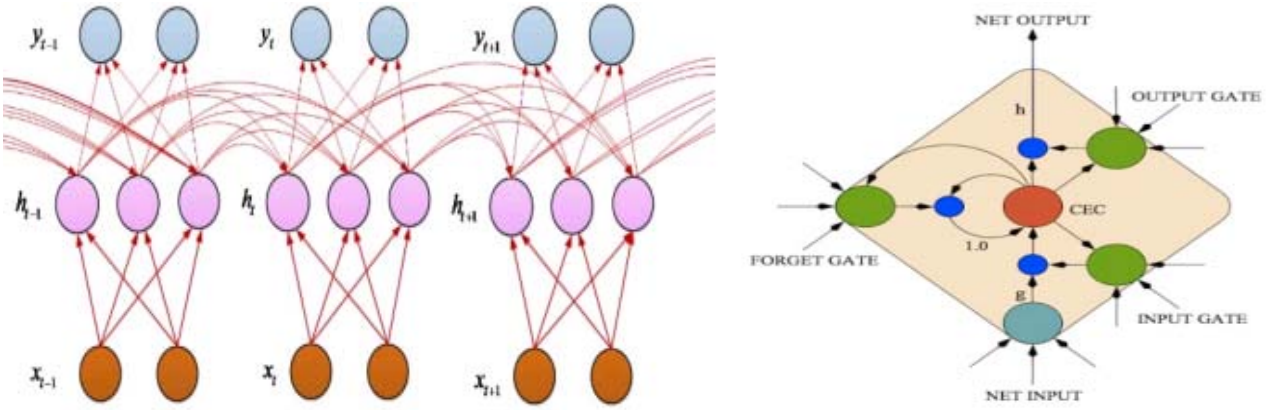
Under the supervision of 5 basic design principles, the thesis has built emotion metaphor corpus which assists model building of machine translation between English and Chinese. Table 2 gives the statistical results of the corpus built in the thesis.

| | Humankind | Living things | Non-living things |
|---|---|---|---|
| Properties in tenors | 0.8761 | 0.0241 | 0.0999 |
| Properties in vehicles | 0.4254 | 0.2325 | 0.3421 |

Table 2. Proportion of Noumenon and Metaphor's in Field's Broad Heading

## 4. Application Comparison of Emotion Metaphor Corpus in Machine Translation

To test and verify the feasibility of the emotion metaphor corpus built in this thesis, the thesis uses tagged data examples and emotion metaphor results based on existing emotion metaphor corpus, adopts deep learning models in machine translation to translate examples, and tests and verifies the relevant emotion metaphor results. During the natural language processing, the ordinary translation is based on the whole sentence. However, actually, words in sentences are closely related, and contextual relations can be found everywhere in both Chinese and English. Therefore, each sentence's machine translation can be seen as a chronological series data. 90% data of every major corpus registered as training sets and 10% as test sets to test tagging results of emotion metaphor in machine translation in both E-C and C-E translation. Error result was reflected by cross entropy [15], and test results and real tagging results' errors served as inverse feedback of deep models. Machine translation process was based




(a) RNN Structure (b)LSTM Mnemon Structure

Figure 2. RNN and LSTM Mnemon Structures

on the time series data, therefore recurrent neural network was adopted in deep models and LSTM mnemon was used to record chronological series relations.

The built-in cyclical structure of Recurrent Neural Network (RNN) ensures chronological relations of series data and stores information. The RNN is shown in Figure 2(a). In this structure, recurrent network modules of RNN deliver information from the previous level to the next, and every output in the network modules' hidden layers depends on the previous input information. Therefore, this network features a cyclical structure. The chained feature of RNN is conducive to handling a whole sentence in machine translation because sentence is made of words and they are chronologically arranged. However, as RNN's layers increase, the related grads will disappear and machine translation may not be effective. Therefore, the thesis adopts LSTM to build RNN with kept grads to achieve equitable machine translation and emotion metaphor prediction results. Figure 2(b) describes the basic structure of LSTM.

LSTM mnemon adopts memory gate, input gate and output gate to ensure data continuity [17], this structure consists of formulae including:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (4)$$

$$c_t = f_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (6)$$

$$h_t = o_t tanh(c_t) \quad (7)$$

In this part, *i, f, o* stand for input gate, memory gate and output gate, *W* for parameters among gates, *σ* for activation functions among gates, *tanh* for activation functions among layers of RNN. By using BPTT, machine translation's emotion metaphor is able to build related recurrent neural network.

Examples in corpus were then denoted. For instance, [Jimmy, B, 11111], [ice ball, Y, 12315] was input in the neural network. After translation, the related emotion metaphor [merry, D], [PA] was used to compare with neural network's expected output. The following formula was used to calculate errors of cross entropy:

$$E = \frac{1}{N}\sum_{t-1}^{T} o_t \log h_t + (1-o_t)\log(1-h_t) \quad (8)$$

In this part, after continuous reverse transfer errors, finally emotion metaphor prediction related with machine translation under the translation between Chinese and English, and the expected results under the training sets were accomplished. In order to test the test set for the trained deep recurrent neural network, 5 existing corpora were trained and tested under the same methods. Results are described in the following Table.

The test results show that the emotion metaphor corpus based on machine translation has strong robustness. With the error-correcting mechanism, professional denotation, and classification of emotion metaphors, the corpus performs at its best level with strong robustness under the deep recurrent neural network. The corpus built in this thesis provides an in-depth reflection of mathematical modeling in emotional metaphor.

| | MML | Metalude | Metabank | XMU Dataset | Our Dataset |
|---|---|---|---|---|---|
| Training set results | 74.28% | 83.29% | 85.38% | 72.38% | 89.32% |
| Test set results | 56.37% | 63.28% | 69.37% | 54.28% | 73.39% |

Table 3. Test and Training Results of 5 Corpora and Deep Recurrent Neural Network

## 5. Conclusions

With the burgeoning development of computer technology, more fields need machine translation. Unlike human, machine lacks emotion metaphors in its translation process, thus may fail to express its real meaning. To promote and improve the machine translation, emotion metaphor's modeling needs to be built in translation. However, the existing corpora are not fit for suitable emotion metaphor models. The thesis has built an emotion metaphor corpus with error-correcting mechanism, clear emotion classification, professional data collection and tagging by designing 5 basic principles. Compared with the existing corpora, the corpus of this thesis shows effective functions and robustness for emotion metaphor modeling and calculation, and provides a strong support for emotional researches of machine translation.

## References


[1] Bahdanau, Dzmitry., Cho, Kyunghyun., Yoshua Bengio. (2014). Neural machine translation by jointly learning to align and translate.

[2] Cho, Kyunghyun, et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation.

[3] Mikolov, Tomas., Quoc V. Le, Sutskever, Ilya . (2013). Exploiting similarities among languages for machine translation.

[4] Shutova., E. (2010). Models of metaphor in NLP. *In*: Proceedings of the 48th Annual Meeting of the Association for Compu-tational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010. 688–697.

[5] Shutova., E., Teufel., S. (2010) Metaphor corpus annotated for source-target domain mappings. *In*: Proceedings of the International Conference on Language Resources and Evaluation, Malta, 3255–3261

[6] Martin., H. (1994). Metabank: a knowledge-base of metaphoric language conventions. *Comput Intel,* 10. 134–149.

[7] Devlin, Jacob., et al. (2014). Fast and Robust Neural Network Joint Models for Statistical Machine Translation. ACL (1).

[8] Irvine, Ann., Callison-Burch, Chris. (2013). Combining bilingual and comparable corpora for low resource machine translation. *In:* Proceedings of the Eighth Workshop on Statistical Machine Translation.

[9] Tian, Liang., et al. (2014). UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. LREC.

[10] Mohammad, Saif., M. (2015). Imagisaurus: An Interactive Visualizer of Valence and Emotion in the Roget's Thesaurus.*In:* 6th workshop on computational approches to subjectivity, sentiment and social media analysis wassa.

[11] Costa-Jussa, Marta R., and José AR Fonollosa. (2015). Latest trends in hybrid machine translation and its applications, *Computer Speech & Language* 32 (1) 3-10.

[12] Och, Franz Josef. (2014). Selection and use of nonstatistical translation components in a statistical machine translation framework. U.S. Patent No. 8,666,725. 4 Mar.

[13] Batjargal, Biligsaikhan, et al. (2013). Applying Text Encoding Initiative Guidelines to a Historical Record in Traditional Mongolian Script. Culture and Computing (Culture Computing), *In*: 2013 International Conference on. IEEE, 2013.

[14] Luong, Minh-Thang, et al. (2014). Addressing the rare word problem in neural machine translation.

[15] Luong, Minh-Thang., Pham,Hieu., Manning, Christopher D.. (2015). Effective approaches to attention-based neural machine translation.

[16] Jan-Thorsten, Peter., Wang, Weiyue., Ney,Hermann. (2016). Exponentially Decaying Bag-of-Words Input Features for Feed-Forward Neural Network in Statistical Machine Translation. *In*: The 54th Annual Meeting of the Association for Computational Linguistics.

[17] Cui, Yiming, et al. (2015). LSTM Neural Reordering Feature for Statistical Machine Translation. arXiv preprint arXiv:1512.00177.