Corpus and dictionary development for classifiers/quantifiers towards French-Japanese machine translation

Mutsuko Tomokiyo
LIG-GETALP, IMAG-CAMPUS
700 avenue Centrale 38401 Grenoble
Mutsuko.Tomokiyo@imag.fr

Christian Boitet
UGA, LIG-GETALP, IMAG-CAMPUS
700 avenue Centrale 38401 Grenoble
Christian.Boitet@imag.fr

Abstract

Although quantifiers/classifiers expressions occur frequently in everyday communications or written documents, there is no description for them in classical bilingual paper dictionaries, nor in machine-readable dictionaries. The paper describes a corpus and dictionary development for quantifiers/classifiers, and their usage in the framework of French-Japanese machine translation (MT). They often cause problems of lexical ambiguity and of set phrase recognition during analysis, in particular for a long-distance language pair like French and Japanese. For the development of a dictionary aiming at ambiguity resolution for expressions including quantifiers and classifiers which may be ambiguous with common nouns, we have annotated our corpus with UWs (interlingual lexemes) of UNL (Universal Networking Language) found on the UNL-jp dictionary. The extraction of potential classifiers/quantifiers from corpus is made by UNLexplorer web service.
Keywords : classifiers, quantifiers, phraseology study, corpus annotation, UNL (Universal Networking Language), UWs dictionary, Tori Bank, French-Japanese machine translation (MT).

1 Introduction

Recent Machine Translation (MT) evaluation tends to be conducted based on (1) Automatic evaluation metrics use reference translations for each segment such as BLEU, NIST, METEOR (Papineni et al., 2001; Banerjee and Lavie, 2005; Doddington, 2002).

This shows frequent efforts for MT, by measuring a similarity or a distance between a translation hypothesis and its post-editions. Basic operations used for post-editions are substitution, deletion, and insertion of words or phrases in a sentence, whatever the MT system is. (2) Subjective measures are based on human judgements of "intelligibility", "fidelity", "adequacy" and "fluency" of MT outputs.

These methods are really suitable for evaluating the progress of MT systems, but they do not contribute directly to improve the quality of MT outputs. Here we focus on lexical ambiguities, which are considered as a main cause of the degradation of the quality in MT for spoken or written sentences. Several types of ambiguity appear on each phase of MT for different types of documents.

We have categorized ambiguity problems according to the levels of MT analysis and to the MT contexts in which they are encountered, and we have proposed a formal ambiguity representation as well as guidelines for ambiguity labelling to build an ambiguity data base[1].

In fact, according to our studies of ambiguities, 14% of analysis errors[2] are due to polysemous words. Also, (G.Wisniewski and al., 2013) say the most frequent necessary post-edition in their French corpus translation into English is to correct articles like «les», «le», «du», etc., and the next one concerns lexical transfer errors of polysemous words. In addition, when polysemous words are used in their abstract or figurative meaning where they could be classifier or quantifier, translation results produced by current

[1]We have done research on ambiguity analysis from the lexical, semantic and contextual points of view since 1996. Ambiguities have been defined, categorized, and formalized as objects in an ambiguity database, and we have used this theoretical background to label ambiguities in Japanese-English interpreted dialogues, collected for the development of a speech translation system at ATR in Japan (1994). (Boitet and Tomokiyo, 1995; Boitet and Tomokiyo, 1996; **?**)

[2]The ambiguity analysis includes assignment of speech acts, although generally speaking speech act ambiguity isn't taken account of, so the percentage is important.

Proceedings of the Workshop on Cognitive Aspects of the Lexicon,
pages 185–192, Osaka, Japan, December 11-17 2016.

MT systems are not at all good. Even measure words like cm, km, kg, etc. may be ambiguous with acronym (Anil K. et al., 2013).
Example: cm → centimètre, congrégation de la mission, coût marginal, etc.
The following example shows that «pincée (pinch, つまみ, tsumami) » in a quantifier phrase appears in form of «une pincée de», and is used in its figurative meaning. When one looks at the translation outputs produced by commercial MT systems, it's not hard to deduce there is a lack of phraseology studies and polysemy disambiguation method for the word «pincée»[3]. For the treatment of the classifier/quantifier expressions, at first, we must know whether a word or an expression in a document is the classifier/quantifier or not, and which kind of information is necessary to handle it in MT.

Example: Ajoutez une pincée de sel. (Add a pinch of salt.) → 塩のつねりを加えなさい/塩のピンチを加えなさい (Shio no tsuneri wo kuwaenasai/Shio no pinchi wo kuwaenasai)[4]

Sections 1 & 2 discuss the problems encountered in the processing of classifiers and quantifiers arising for meaning determination in the source language and from the structural differences between a language pair in the framework of MT. Section 3 describes morpho-syntactic problems between two languages for quantifier/classifier expressions. In Section 4, the difficulty of quantifiers/classifiers extraction is described. In Section 5, we propose a solution using a dictionary, edited from collected documents, themselves annotated with semantic UNL (hyper)graphs, presented as a parallel corpus, and give somme details about a small French-Japanese dictionary for quantifiers/classifiers, built for MT experimentation with an UNL system[5].

2 Lexical ambiguity for classifiers/quantifiers

We call here words or phrases which are used in some languages to indicate the class of nouns or nominal/adjectival phrases, depending on the type of these referent, classifiers/quantifiers, when they appear in quantitative expressions. They denote:
(a) temporal/spatial quantity of the referent and
(b) states of the referent in an idiomatic expression.
Type (a) classifiers/quantifiers express concrete measurement, and type (b) classifiers/quantifiers express quantitative states of the referent based on speaker's observation.

Examples:
Type (a): 2g de sel (2グラムの塩, 2-guramu-no shio, 2g of salt)
Type (b): une pièce de viande (一切れの肉, hitokire-no niku, a piece of meat) / un brin de causette (ちょっとしたおしゃべり, chottoshita osyaberi, a little chat)

Classifiers/quantifiers of type (a) are obligatory in quantitative expressions, and they often cause acronym ambiguities for MT as mentioned above, and also ambiguities due to the "floating quantifier" (Inoue, 1989) phenomenon in Japanese.
For classifiers/quantifiers of type (b), there are three different sorts of problems. The first one is the fact that classifiers/quantifiers have many to many meaning corespondences between source-target languages pairs. In the following example, the French word «pièce» is translated into «切れ, kire», «枚, mai», «点, ten», «頭, tou», etc. in Japanese, because, in many cases, Japanese classifiers depend upon the visual forms of referents.

The second problem arises in the case where classifiers/quantifiers don't appear explicitly in one language of a language pair, nevertheless they are mandatorily expressed in the other, like «冊», satsu in Japanese.

[3] "pincée" is used as quantifier/classifier for pulverized substances.
[4] These translations don't make sense. http://www.reverso.net/translationresults.aspx?langFR&directionfrancais-japonais. http://www.worldlingo.com/fr/products_services/worldlingo_translator.html.
[5] The UNL (Universal Networking Language) system denotes a language for computer, multilingual encoder-decoder system, UNL-UWs dictionary, parallel corpus, and linguistic ontology system. It has been developed under the aegis the Organization of United Nations University in form of international consortium for written languages processing since 1996. We are one of the pioneer members of the consortium. Bilingual dictionaries with UNL-UWs dictionary are edited by each "UNL language center". http://www.undl.org/unlsys/unl/unl2005/attribute.htm

Table 1: Translation of French word ”pièce” into Japanese

French entries	Examples	Source	Japanese translations
pièce	une pièce de toile	Royal	一枚(mai)の布 (ichimai no nuno, a piece of cloth)
	une pièce de mobilier	Royal	一点(ten)の家具 (itten no kagu, a piece of furniture)
	dix pièces de bétail	Royal	１０頭(tou)の家畜 (jyuttou no kachiku, ten cattles)
	plusieurs pièces de bois	Royal	数枚(mai)の板 (suumai no ita, some boards)
	une pièce de vin est un tonneau de vin contenant environ 220 litres.	Wiki, pièce	一樽(hitotaru)のワインとは 約２２０リットルを含むワイン 樽である (hitoraru no wain toha yaku 220 littoru wo fukumu waindaru dearu, a barrel of wine includes 220 littles of wine)
	J’ai reçu une demi-pièce de ce vin.	Vinothèque	わたしは半樽(hantaru) のワインを受け取った。 (watashiha hantaru no wain wo uketotta, I have received half barrel of this wine.)
	Dans une pièce de théâtre, il n’y a pas de narrateur pour raconter les faits.	http://www.etudes-litteraires.com/etudier-piece-de-theatre.php	ある作品(sakuhin) では事実を語るナレータがいない。 (aru sakuhin deha jijitsuwo kataru nare)ta) ga inai, There is no narrator in a program.)
	Une pièce de viande	Royal	一切れ(kire)の肉 (hitokire no niku, a slice of meat)
	Une pièce de blé	Royal	一枚 (mai)の麦畑 (ichimai no mugibatake, a field of wheat)

Table 2: Translation of the French expression «pointe» into Japanese

French entries	Examples	Source	Jp translation	E.n translation
Pointe	une pointe d'ironie mal placée	J.L. Carré	場違いの皮肉をちくりと	the tip of , a hint of, a note of, a trace of
	relever la sauce avec une pointe d'ail	Livre de cuisine	ソースにニンニクをちょっときかせる	pick up the sauce with a hint of garlic
	avec une pointe d'agacement dans la voix	T. Jonquet	声にすこし苦しみをにじませて	with a hint of irritation in the voice
	mettre une pointe d'ironie dans sa question	Royal	質問にちくりと皮肉を込める	with a suggestion of sarcasm

Examples:
2 livres → 二冊の本 (ni-satsu no hon, two books)
un chat → 一匹の猫 (i-ppiki no neko, a cat) (see →Table 1)

The third problem occurs during the analysis/transfer phase as locutions problem like «un brin de»: «brin» signifies «茎, kuki, small stalk», and «un brin de» means «a little of». It's translated into «ちょっとした (chottosita, small)» in Japanese. This is due to the polysemy of «brin» and to the cognitive or metonymic differences between two languages.

Table 3: KWIC of “pointe” from Sketch Engine

doc#357	qui marque le déclin définitif de cette	pointe	de poussée et de sécrétions des hormones
doc#397	la sierra Pacaraima, qui constituent une	pointe	avancée du Sertao brésilien. En janvier
doc#457	de nouveauté, un soupçon de douceur, une	pointe	d’exotisme : commence par te mettre dans
doc#517	Tafer ne sont capables d’évoluer seuls en	pointe	. Arles - Marseille En concédant une

3 Morpho-syntactic differences between French-Japanese classifiers/quantifiers

As for the behaviour of floating quantifiers in Japanese (Inoue K.1989), the problem we encounter in building a Japanese-French MT lies in the fact that the Japanese quantifiers can be freely positioned between phrasal units in a sentence except after predicative verbs. They are morphosyntactically classified into two types of quantifier expressions: (1) noun phrases in form of “Number+Quantifier+の(no, of)+Noun («NQのN» type)”, and (2) noun phrases in form of Noun+Number+Quantifier («NQN» type).
The NQN type can syntactically be divided into «N» part and «QN» part and it’s possible to use «QN» like an adverb before a predicative verb in a sentence.
Hence, three types of expressions are possible for the same meaning : (1) 二冊の本 (ni-satsu no hon, two books), (2) 本二冊 (hon ni-satsu, two books) and also (3) 本を二冊 (hon-wo ni-satsu, two books)[6]. The floating quantifier can produce meaningless translation result in some cases. For instance, “3kgの子豚がいました (3 kiloguramu no kobutaga imashita, There was a 3kg piglet.)” is acceptable as Japanese sentence, but “子豚が3kgいました(kobuta ga 3kiloguramu imashita)”[7] doesn’t literally make sense, because, «子豚 (kobuta, piglet)» means only an alive pig and co-occurs with いました (there was), but “3kg” cannot do [12]. So, to avoid the translation output “子豚が3kgいました”, we need to have supplementary information for “子豚 ” and the verb “いる(iru, there is, or exist)” and implement method to use it. For that reason, we use the UWs dictionaries of the UNL system, which allows us to describe semantic constraints between words.

4 Recognition difficulty of quantifiers/classifiers

We extract type (a) quantifiers/classifiers from Tori Bank[8](See Annex), while referring to existing weights and measures dictionaries. For type (b) quantifiers/classifiers, it’s laborious to pin down phrasemes[9] in row data.
Eg. “pointe” from Sketch Engine (Table 3).

However, French and English phrasemes are, in many cases, composed of “Number+Noun+de (Number+noun+of)+Noun without particle” like “une poignée de sable (a handful of sand)”, “une pointe d’ironie (a touch of irony)”, “un pouce de terre (a handful of)”, so in order to collect data of type (b), we take note of the morphologic characteristic (Petit, 2004), and utilize a multilingual corpus management software, called Sketch Engine[10]. The software gives a list of tri-grams of keywords in context. The used documents are journals, magazines, novels, existing expression dictionaries, French, Japanese and English leaner’s manuals. The assignment of the QC for obtained keywords is made by linguistic intuition, while watching output from MT experiences on UNL Explorer[11].

[6]In the full sentence, I bought 2 books (本を二冊 買いました, hon-wo nisatsu kaimashita)
[7]子豚が3kgいました, For the piglet, there was 3 kg*.
[8]Tori Bank is a phrase corpus which has been developed at Tottori Unversity in Japan in 2007. http://unicorn.ike.tottori-u.ac.jp/toribank/about_toribank.html
[9]The term ”phraseme” means set phrase, idiomatic phrase, polylexical expression, etc.
[10]The Sketch Engine refers to a text corpus management and analysis software developed by Lexical Computing Limited since 2003. (http://en.wikipedia.org/wiki/Sketch_Engine)
[11]UNL Explorer is a web-based application, which combines all the components of the UNL system to be accessible online.

5 Specification of classifiers/quantifiers corpus

The corpus includes sentences which are manually or semi-automatically collected from novels, cooking articles, news papers, dictionaries, Tori Bank, etc. The description for ”pointe” is given below as a typical example. The annotated keywords include, at the present time, about 1000 classifier/quantifier expressions for Japanese, French and English in PhraseBook II[12] (see Annex).

1. Identification number: XX
2. Keywords and class: pointe (n.)
3. English sentence: to season the sauce with a hint of garlic
4. French sentence: relever la sauce avec une pointe d’ail
5. Japanese sentence: ソースにニンニクをちょっときかせる
6. Source: Royal
7. UNL annotation (simplified):

{org:fr} Relever la sauce avec une pointe d’ail {/org}
{unl}
agt(season(agt>person, obj>dish, icl>action>thing).@entry.@imperative, you)
obj(season(agt>person, obj>dish, icl>action>thing).@entry.@imperative, sauce(icl>cooking).@def)
met(season(agt>person, obj>dish, icl>action>thing).@entry.@imperative, garlic(icl>cooking))
qua(garlic(icl>cooking), a hint of(icl>quantity))
{/unl}
{en} Season the sauce with a hint of garlic {/en}
{jp} ソースにニンニクをちょっときかせる {jp}

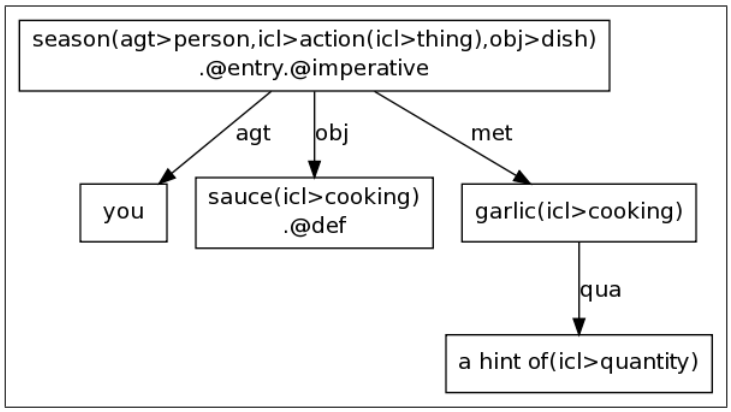

Figure 1: UNL graph for ”Relever la sauce avec une pointe d’ail”

6 UNL-UWs dictionary for quantifiers/classifiers

The collected documents in Japanese, French, English are annotated by their UNL expressions[13], which are composed of interlingual lexemes called “universal words (UWs)[14], semantic boolean features, ” and semantic relation tags[15]. In general, a UW is made of an English word or locution, its ”headword”, disambiguated by a list of restrictions. The set of UWs can be used as a lexical “pivot” between the

[12]The corpus is going to become larger by extracting classifiers/quantifiers expressions from Tori Bank

[13]UNL is a language for computer to represent the meaning of natural language expressions. The ”Universal Words” (UWs) constitute its vocabulary. A UW is in effect an*interlingual lexeme*. Each node of a ”UNL expression” (in effect, a semantic hypergraph) bears a UW and a possibly empty set of semantic attributes (Uchida et al.,2006).

[14]The UNL-UWs dictionary contains, at the moment 1269421 word senses (mapped to as many UWs) for Japanese, 520305 word senses for French, and 1458686 word senses for English.

[15]The semantic relations are represented by a fixed set of 42 relation 3-letter symbols, like agt, aoj, gol, etc., and the attributes are boolean, like .@def or .@soon-begin. There are about 200 attributes in the UNL specifications, and developers may introduce new attributes. These predefined attributes include syntactic, semantic or pragmatic information. The annotation labels are in fact, “icl”, “equ”, “quantity”, etc. in description example.

"lexical spaces" of any set of natural languages, and the UNL graphs can similarly be used as an "anglo-semantic" abstract pivot language. The added information for classifier/quantifier expressions is merged into the UNL dictionaries. Here is an extract of our 3-lingual UNL dictionary. The first entry has 2 languages (jp, fr.). The second entry has 3 languages (jp, fr, en). The forth has again 2 languages (fr, en).
樽 (taru, pièce): cask(icl>wine,equ>2200 litres)
冊 (satsu, volume): volume(icl>quantity)
relever (to season): season(agt>person, obj>dish, icl>action>thing)
pointe (touch): touch(icl>amount) → une pointe de (a touch of)

"icl" and "equ" in our UW dictionary are semantic relation tags, and mean headword's sub-meaning and equivalent quantity, respectively. The semantic relation "agt" indicates that the volitional agent of "relever" is "person".

Perspectives and Conclusion

We are making French-Japanese MT experiments using the UNL system.
We have studied the methodology for phraseology treatment on MT systems while developing a French-Japanese-English parallel corpus and have concluded that a deeper linguistic analysis (Petit, 2004, Mari, 2011) is necessary for UW dictionary description. Our corpus will be useful for software developers, as well as for learners of languages, because it covers semantic information which cannot be yet found in any bilingual dictionary. We also plan to develop a language software by processing the corpus, where corresponding words between 2 languages are shown on demand by character blinking or where the meaning of nouns or verbs in a sentence is shown without any ambiguity by interpreting the UNL annotations. A prototype of the software has been already presented in a PhD thesis (Chenon, 2005).

References

S. Banerjee and A. Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. Association of Computational Linguistics (ACL), Association of Computational Linguistics (ACL).

Christian Boitet and Mutsuko Tomokiyo. 1995. Ambiguity and ambiguity labelling : towards ambiguity data bases. In Proc. of RANLP-95, Bulgaria. Recent Advances in Natural Language Processing, Recent Advances in Natural Language Processing.

Christian Boitet and Mutsuko Tomokiyo. 1996. On the formal definition of ambiguity and related concepts, leading to an ambiguity-labelling scheme. In Actes de MIDDIM-96. MIDDIM-96 Post-COLING Seminar, GETA.

Christophe Chenon. 2005. Vers une meilleure utilisabilité des mémoires de traductions, fondée sur un alignement sous-phrastique. Ph.D. thesis, Université Joseph Fourier, Grenoble.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proc. of HLT-2002, pages 128–132, San Diego. Human Language Technology, Human Language Technology.

Céline Gouverneur. 2005. The phraseological patterns of high-frequency verbs in advanced english for general purposes. TaLC 6.

Kazuko Inoue. pages 201–204. Taisyûkan-syoten.

Alda Mari. 2011. Quantificateurs polysémiques. Ph.D. thesis, Université Paris-Sorbonne.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Research report RC22176, IBM.

Gérard Petit. 2004. La polysémie des séquences polylexicales, syntaxe et sémantique. HAL, (Id-00648029).

Mutsuko Tomokiyo and Monique Axtmyer. 1996. Experiments in ambiguity labelling of dialogue transcriptions. MIDDIM-96 Post-COLING Seminar, MIDDIM-96.

Hiroshi Uchida, Meiyin Zhu, and Tarcisio G. Della Senta. 2006. Universal Networking Language. UNDL Foundation, Japan.

Guillaume Wisniewski, Anil Kumar Singhand Natalia Segala, and François Yvon. 2013. Un corpus d'erreurs de traduction. Les Sables d'Olonne, France. TALN-RÉCITAL, TALN-RÉCITAL.

Annex 「鳥バンク」(Tori Bank)

Examples: 「 塁 (rui, base)」 , 「寸 (sun, approx. 3.03 cm)」

AC00046100 P 11:二塁走者の生還を許し :VP@28:allowing the runner to score from second:VP

AC00046100 P 4:一塁へ悪投し、:VP@7:threw wild to first:VP

AC01599600 C6:一寸先も見え:CL@27:we could not see an inch ahead:CL

AC01599600 P6:一寸先も見え:VP@40:see an inch ahead:VP