A contrastive review of paraphrase acquisition techniques

Houda Bouamor, Aurélien Max, Gabriel Illouz, Anne Vilnat

LIMSI-CNRS & Univ. Paris Sud
Orsay, France
firstname.lastname@limsi.fr

Abstract

This paper addresses the issue of what approach should be used for building a corpus of sentential paraphrases depending on one's requirements. Six strategies are studied: (1) multiple translations into a single language from another language; (2) multiple translations into a single language from different other languages; (3) multiple descriptions of short videos; (4) multiple subtitles for the same language; (5) headlines for similar news articles; and (6) sub-sentential paraphrasing in the context of a Web-based game. We report results on French for 50 paraphrase pairs collected for all these strategies, where corpora were manually aligned at the finest possible level to define oracle performance in terms of accessible sub-sentential paraphrases. The differences observed will be used as criteria for motivating the choice of a given approach before attempting to build a new paraphrase corpus.

Keywords: Paraphrase acquisition, Corpus collection, Paraphrase typology

1. Introduction

Paraphrases constitute admittedly one of the basic types of knowledge required by Natural Language Processing systems, and automatic acquisition has attracted a lot of work (Madnani and Dorr, 2010). One of the most common approaches for acquiring *sub-sentential* paraphrases is to first collect pairs of sentences. If these pairs are paraphrases, then it is expected that they contain many precise sub-sentential paraphrases. Otherwise, the nature of the sentence pairs will have an impact on the numbers and types of sub-sentential paraphrases that they contain. In this article, we address this issue on French by considering a large range of possible strategies for collecting sentence pairs, and performing an annotation study to describe the sub-sentential paraphrases that are found in each corpus type. The objective of this paper can be paraphrased as that of providing information for selecting a type of corpus depending on specific requirements for paraphrase acquisition.

We will first briefly review the domain of paraphrase corpus construction (section 2.). We will then describe the 6 strategies that we have studied for this work (section 3.): (1) multiple translations into a single language from another language; (2) multiple translations into a single language from different other languages; (3) multiple descriptions of short videos; (4) multiple subtitles for the same language; (5) headlines for similar news articles; and (6) sub-sentential paraphrasing in the context of a Web-based game. An experiment were 50 sentence pairs for each corpus types were manually annotated will then be reported (section 4.1.). Annotation was done at the finest possible level to define oracle performance in terms of accessible sub-sentential paraphrases. We will describe a typology of sub-sentential paraphrases per corpus type (section 4.2.) and conclude (section 5.). The differences observed may be used in future research as criteria for motivating the choice of a given approach before attempting to build a new paraphrase corpus.

2. Building paraphrase corpora

Finding equivalent text units at different levels (lexical, phrasal or sentential) can be beneficial to a number of NLP tasks, such as Information Retrieval (Pasça, 2005), Machine Translation (Schroeder et al., 2009), and Summarization (Hirao et al., 2004). We briefly review here the main approaches for building paraphrase corpora.

Several works have used monolingual parallel corpora obtained by multiple human translations of the same foreign source text. For instance, Barzilay and McKeown (2001) exploited monolingual parallel corpus of multiply-translated novels and used contextual information based on lexical similarity to extract paraphrases. The corpora described by Cohn et al. (2008) include sentential paraphrases acquired from multiply-translated Chinese news sentences into English.

Because this type of resource is extremely scarce, techniques based on monolingual comparable corpora have been proposed, as this type of resource is much more readily available. Such corpora often correspond to texts grouped with respect to common subject and time frame. Barzilay and Elhadad (2003) learned rules that exploit topic structure and local alignment to extract sentence pairs from news stories about the same events from different press agencies. Dolan and Brockett (2005) built the Microsoft Research Paraphrase Corpus, which contains sentence pairs collected using heuristic extraction techniques and a SVM classifier to select likely sentential paraphrases from a large corpus of thematically clustered news texts. An in-depth review of these methods is given in (Madnani and Dorr, 2010).

Provided corpora are large enough, paraphrase acquisition techniques can potentially extract a large number of sub-sentential paraphrases. However, corpora may have genres and/or domains that are not adapted to specific requirements. Furthermore, it may be the case that some useful textual units are never observed in such corpora. Collaborative resource creation has therefore been used to overcome in part those difficulties. Paraphrase acquisition has

for instance been formulated in the context of a Web-based game (Chklovski, 2005), or as independent descriptions of short videos through crowdsourcing (Chen and Dolan, 2011). Other works have also considered looking at histories of collaboratively authored text such as the Wikipedia encyclopedia to identify rewritings corresponding to paraphrases (Max and Wisniewski, 2010).

Lastly, the field of automatic paraphrase generation (Madnani and Dorr, 2010) has produced numerous techniques that have been used for generating sentential paraphrases used for learning Statistical Machine Translation systems (Nakov, 2008) or optimizing their parameters (Madnani et al., 2008), and various potential uses including uses by humans (Quirk et al., 2004; Zhao et al., 2010). In addition to the fact that the current performance of these techniques is limited, it should be noted that an issue of circular dependency arises, as these techniques rely on the availability of paraphrase resources; most of these techniques are, however, able to produce paraphrases that did not belong to their original resources.

3. Corpus types

We now describe the various corpus types that we have considered in our study. We have selected 6 strategies for collecting sentential paraphrase pairs, which, as our analyses will show, correspond to various degrees of sentence parallelism, and therefore allow to extract different numbers of sub-sentential paraphrases. Examples for each collected corpus types appear in Table 1.[1]

3.1. Multiple translations from a single language or from different languages

Multiple independent human translations of the same texts produce sentential paraphrases (Mitsuo Shimohata and Matsumoto, 2004), which can subsequently be used for sub-sentential paraphrase acquisition (Barzilay and McKeown, 2001). We built two sub-corpora from the MULTITRAD corpus (Bouamor, 2010), a paraphrase corpus for French built by collecting multiple translations from several languages proposed by volunteers. We selected a set of 50 sentences on the basis that 4 independent valid translations from English and one valid translation from German, Spanish, Italian and Portuguese were available. In each group, one paraphrase was randomly selected as a "reference paraphrase" which was paired to the three others. Two sub-corpora were built, one from a single language (English), and another one from different languages (German, Spanish, Italian and Portuguese). The first two rows of Table 1 illustrate examples from those two corpora. The last line in *italics* is the reference sentence in English, which has been translated twice into French. This same sentence, which was also available in German, Spanish, Italian and Portuguese, was also translated into French as shown in the examples of the second row of Table 1.

3.2. Multiple translations of subtitles

Many subtitles for movies and TV series are contributed by Web users in many languages, and several users sometimes

independently contribute subtitles for the same video in the same language. We used the sentence alignment procedure that exploits time frames described in (Tiedemann, 2007), and collected a monolingual corpus for 50 sentence pairs of two subtitles of the TV series *Desperate housewifes*. The third row of Table 1 illustrates a sentence uttered in English (in *italics*) and its two proposed subtitles in French.

3.3. Multiple descriptions of Videos

Another type of semantic content for acquiring paraphrases is the scene recorded on a video. Chen and Dolan (2011) used crowdsourcing to acquire multiple independant descriptions of short videos from contributors speaking different languages. We extracted from this Microsoft Video Description Corpus 50 description pairs in French. Note, however, that these descriptions did not have the "verified" status indicating that their contributors were manually selected. In Table 1, we give an example of two descriptions in French of a same short video, with one corresponding description in English (in *italics*) from the corpus.

3.4. Clustered news headlines

News article headlines can be regarded as a potentially rich source of paraphrases, as a single event may be reported in diverse ways (Wubben et al., 2009). News headlines can be easily collected from the Web, for example by using a news aggregation service. We used headlines from the French version of Google News[2] clusters: we initially retrieved 100 clusters from the service, and kept the 50 headline pairs with minimal edit cost as given by TER (Snover et al., 2006). Table 1 gives an example of such a pair, with a possible translation into English.

3.5. Web-based paraphrasing game

Paraphrase acquisition has also been formulated as a language game (Chklovski, 2005). We have designed and implemented a Web-based game for sub-sentential paraphrase acquisition, where players have to propose paraphrases in some predetermined context. The rules of the game give bonus points to players who propose paraphrases that receive high marks from other players (who do not take part in the same games) and are rare relative to the set of proposed paraphrases. We have collected paraphrases from 22 French-speaking volunteers who contributed paraphrases for 20 sentential paraphrase pairs. Half of them were given the first sentence of a sentential paraphrase pair for context, and had to paraphrase the appropriate phrase, while the other half were given the second sentence and the corresponding phrase to paraphrase. An example is given in the last row of Table 1.

4. Annotations and results

We now analyze the results of an annotation experiments on French (section 4.1.) and detail a typology of sub-sentential paraphrases that can be acquired from all corpora under study (section 4.2.).

---

[1]Note that the examples are given as they appear in the different corpora, which explain some typos and other errors.

Source	Paraphrases
SINGLE LANGUAGE TRANSLATION	→ Plusieurs orateurs ont considéré que ceci est trop tardé. → Plusieurs locuteurs ont jugé cela nécéssaire depuis longtemps. • *Several speakers considered this to be long overdue.*
MULTIPLE LANGUAGE TRANSLATION	→ Plusieurs intervenants l'ont considéré comme une chose indispensable. → Le retard avec lequel s'accomplie cette étape a été souligné dans de nombreuses interventions. • *Several speakers considered this to be long overdue.*
MULTIPLY-TRANSLATED SUBTITLES	→ On ne voudrait pas qu'ils imaginaient qu'on n'est pas heureux → Personne ne doit douter de notre bonheur conjugal • *No one should suspect our marital bliss*
VIDEO DESCRIPTION	→ Superman déplace des rochers. → Superman dégage l'entrée d'une grotte bloquée par des rochers. • *Superman moves the rocks.*
NEWS HEADLINES COLLECTION	→ Algues vertes: un décret favorisera leur prolifération. → Algues vertes: parution d'un décret controversé sur l'épandage. • *Green algae: controversial decree on spreading*
PARAPHRASING GAME	→ La Commission adressera directement à l'Honorable Membre et au Secrétariat du Parlement les informations **dont elle dispose** → La Commission adressera directement à l'Honorable Membre et au Secrétariat du Parlement les informations **en sa possession** • *The Commission should address directly to the Honourable Member and to Parliament's Secretariat information* ***in its possession***

Table 1: Examples of acquired paraphrases from different sources

4.1. Annotation experiment and analysis

Table 2 provides various statistics for the corpora that were built. The first observation is that MULTIPLE LANGUAGE TRANSLATION and SINGLE LANGUAGE TRANSLATION contain significantly larger sentences (resp. 25.5 and 21.4 tokens on average), almost twice longer than those of SUBTITLES. VIDEO DESCRIPTIONS contain very short sentences. The case of PARAPHRASING GAME is different as we only take into account the phrases which have been paraphrased, so we only report the number of tokens for those phrases. To compare the similarity of the sentence pairs, we compute different scores, including various metrics designed to evaluate machine translation output:

- an *Overlap Coefficient (OC)*, which represents the percentage of lexical overlap between the vocabularies of the sentence pairs (computed on lemmas);

- the BLEU score (Papineni et al., 2002), based on *n*-gram precision;

- the TER score (Snover et al., 2006) based on edit rate;

- the METEOR score (Lavie and Agarwal, 2007), based on an harmonic mean of precision and recall[3].

SUBTITLES is the type of corpus with the lowest similarity values. MULTIPLE LANGUAGE TRANSLATION produces, not surprisingly, paraphrases that are less similar than those produced by SINGLE LANGUAGE TRANSLATIONS. Overall, this latter corpus type contains the most similar paraphrase pairs.

We also conducted experiments to study the coverage of *gold standard* sub-sentential paraphrases. To this end, we set up annotation work to annotate the sub-sentential paraphrases in the sentence pairs of the different corpora. Annotation was performed by following most of the guidelines from (Cohn et al., 2008)[4] using the YAWAT Web-based tool (Germann, 2008), . The main guidelines are that *sure* and *possible* paraphrases must be distinguished, smaller alignments are to be prefered but any-to-any alignments may be used, and sentences should be aligned as much as possible. Henceforth, we will only consider for all reported statistics and experiments those paraphrases that are not identity pairs (e.g. (*nice umbrella* ↔ *nice umbrella*)), as they may be considered as trivial as far as acquisition is concerned.

Table 3 reports inter-annotator agreement[5] values computed on sets of 50 sentence pairs. We observe that acceptable values are obtained for sure paraphrases, but that low values are obtained for possible paraphrases. This was somehow expected, given the many possible interpretations of possible paraphrases.

Table 3 also gives proportions and absolute numbers of paraphrases of each type for all corpora. We find that there are globally quite the same number of sure and possible paraphrases in the union of all corpora (resp. 1,846 sure and 1,755 possible). Other salient results include the fact that

---

[3]Note that we do not take synonymy into account here as WordNet synonyms are not available for French.

[4]See `http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_guidelines.pdf`.
Note that in our experiments initial alignments were not obtained automatically.

[5]For each paraphrase type, we used the average of recall values obtained for each annotator set as the reference.

	Corpus statistics			Similarity			
	# tokens	# different tokens	# tokens per sent.	OC(Lemma)	TER	BLEU	METEOR
SINGLE LANGUAGE TRANSLATION	2,138	320	21.4	73.3	56.7	28.2	55.1
MULTIPLE LANGUAGE TRANSLATION	2,594	494	25.5	71.9	70.6	19.3	42.7
MULTIPLY-translated SUBTITLES	1,426	520	14.3	66.3	89.2	14.0	31.1
VIDEO DESCRIPTION	676	70	6.8	72.2	54.7	20.1	39.3
NEWS headlines COLLECTION	925	441	9.2	67.2	52.6	31.8	52.2
Paraphrasing Game	*519*	*160*	*2.6*	*29.6*	-	-	-

Table 2: Statistics on all our corpora (50 sentences) with sentence pair average similarities using several metrics: Overlap Coefficient, TER, BLEU and METEOR

the two TRANSLATION corpora contain roughly the same number of tokens in paraphrases (1,234 vs. 1,220), but that the proportion of *sure* and *possible* is inversed: the paraphrases found in translations from the same language appear to be much more certain to our annotators than those obtained from multiple languages, which is certainly a result of translation sequences. SINGLE LANGUAGE TRANSLATION contains significantly more paraphrases than the other corpora, followed by the SUBTITLES corpus. At the other extreme, the VIDEO DESCRIPTION corpus contains significantly fewer paraphrases, both in proportion and number.

While the translation-based methods facilitate the identification of paraphrases, especially when they are obtained from a single language, such corpora are of limited availability since multiple translations on a large scale are not available. Furthermore, there are not available for any domain, and their construction is time-consuming and requires skilled translators. News headlines are on the other hand easy to acquire, but they are from a limited genre, which implies that extracted paraphrases could be of limited applicability on other genres and domains.

4.2. Typology of sub-sentential paraphrases per corpus type

After the quantitative study of section 4.1., we now consider the nature of the obtained sub-sentential paraphrases. We first define different types of paraphrases, and compare the results obtained for all our corpora.

The different classes that we have asked our annotators to annotate were:

- agreement variations: this class includes number and gender variations for nominal or adjectival phrases, and tense modifications of verbs, e.g. *souhaites ↔ souhaite* (*wish ↔ wishes*);

- inclusion: cases where one of the phrases is more precise than the other, e.g. *droits qu'ils ont acquis ↔ droits acquis* (*rights that they acquired ↔ acquired rights*);

- typographical variations: includes all variations in number (letters or digits), as well as the use of acronyms vs. complete names, e.g. *UE ↔ Union Européenne* (*EU ↔ European Union*);

- morphological variations: e.g. *en Chine ↔ chinois* (*in China ↔ chinese*);

- syntactic variations: where the paraphrases correspond to different syntactic construction, e.g. *Souvenons-nous ↔ On se rappelle* (*Let us recall ↔ We remember*);

- synonymy: different cases of lexical and phrasal equivalence, e.g. *Vous allez adorer ça ↔ ça va vous plaire* (*you are going to love it*);

- pragmatic variations: variations which may only be recognized as paraphrases in the context in which they are encountered, e.g. *manifestants ↔ écologistes* (*protesters ↔ green activists*), *au Bangladesh ↔ dans ce pays* (*in Bangladesh ↔ in this country*).

Table 4 reports the results we obtained. We observe that *Synonymy* (lexical or phrasal) is the most represented category for all corpus type. However, this category only represents 28.3% for NEWS HEADLINES, significantly less than in the other corpora. We further see that the NEWS HEADLINES corpus contains balanced proportions of all types of paraphrases, while, for instance, the *Synonymy* and *Agreement* categories account for three quarters of all paraphrases in the SIMPLE LANGUAGE TRANSLATION corpus.[6] Comparing the two translated corpora, we observe that the main difference is due to the number of inclusion which is more important for translations from multiple languages. The GAME corpus is quite different from the other corpora. Players, by following the rules of the game, proposed mainly synonymy or syntactic constructions, and, for instance, did not considered typographical variations as interesting. If this corpus contais a similar proportion of synonymy than the other corpora, we note that it contains significantly more syntactic variations than any other corpus.

5. Conclusions

In this article, we have proposed a contrastive review of the types and numbers of sub-sentential paraphrases that can be found in various types of sentence pair corpora in French. We have observed that the most parallel sentence pairs in our study were obtained with SINGLE LANGUAGE TRANSLATION, and that the most dissimilar are from VIDEO DESCRIPTIONS. Not surprisingly, we have found that the most similar sentence pairs contain the most sub-sentential paraphrases, especially *sure* paraphrases. However, this type of

---

[6]Note that the important number of typographical variations in the NEWS HEADLINES corpus is probably due to the important amount of acronyms.

	Annotator agreements		Tokens in paraphrase statistics (not considering identity paraphrases)			
			sure para.		possible para.	
	sure para.	possible para.	% tokens	# tokens	% tokens	# tokens
SINGLE LANGUAGE TRANSLATION	64.6	16.6	40.46	865	17.26	369
MULTIPLE LANGUAGE TRANSLATION	61.2	13.4	19.01	493	28.03	727
MULTIPLY-TRANSLATED SUBTITLES	82.7	20.8	23.50	335	35.13	501
VIDEO DESCRIPTION	42.8	9.3	8.14	55	4.73	32
NEWS HEADLINES COLLECTION	67.8	3.8	10.59	98	13.62	126

Table 3: Statistics on the human annotation of subsentential paraphrases recognized in sentence pairs from all corpora divided in *sure* and *possible*.

	Agreement	Inclusion	Typo.	Morpho.	Synt.	Syno.	Pragma
SINGLE LANGUAGE TRANSLATION	28.5	2.1	9.0	3.0	6.6	46.9	3.6
MULTIPLE LANGUAGE TRANSLATION	20.7	20.7	9.3	0.7	4.3	43.6	0.7
MULTIPLY-TRANSLATED SUBTITLES	33.8	8.9	5.3	0.0	5.3	46.4	0.0
VIDEO DESCRIPTION	14.2	8.0	14.2	3.5	11.6	45.5	2.6
NEWS HEADLINES COLLECTION	12.2	16.0	19.7	7.4	8.6	28.3	7.4
PARAPHRASING GAME	0.0	2.0	0.0	4.0	40.0	50.0	4.0

Table 4: Distribution of paraphrase categories in 50 randomly selected sentence pairs from each corpus

corpus is obviously difficult to obtain or costly to build. On the other hand, MULTIPLY-TRANSLATED SUBTITLES and NEWS HEADLINES can be easily collected, on a daily basis for the latter type. These corpora contain sentence pairs that are quite similar, and contain interesting numbers of sub-sentential paraphrases, even if they contain slightly more *possible* than *sure* paraphrases. We have also noted that the NEWS HEADLINES corpus type contains paraphrases of all types. Lastly, we have shown that a corpus from a PARAPHRASING GAME could be used to obtain useful sub-sentential paraphrases.

Figure 1 provides a subjective account of the characteristics of all the studied corpus types which may be useful to select an appropriate corpus for paraphrasing studies or paraphrase acquisition. The characteristics that were considered are the following:

- *Type of corpora*: available or yet-to-be-built;

- *Domain*: domains covered by the acquired paraphrases;

- *Languages*: languages in which corpora can be available or built;

- *Parallelism degree*: degree of meaning preservation;

- *% of aligned paraphrases*: density of sub-sentential paraphrases;

- *Ease of construction*: acquisition cost, time required, etc.

6. References

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of EMNLP*, Sapporo, Japan.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Actes de ACL*, Toulouse, France.

Houda Bouamor. 2010. Construction d'un corpus de paraphrases d'énoncés par traduction multilingue multisource. In *Récital-TALN*, Montréal, Canada.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of ACL-HLT*, Portland, Oregon, USA.

Timothy Chklovski. 2005. Collecting paraphrase corpora from volunteer contributors. In *Proceedings of the 3rd international conference on Knowledge capture*, Banff, Canada.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for development and evaluation of paraphrase systems. *Computational Lingustics*, 34(4):597–614.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, Jeju Island, South Korea.

Ulrich Germann. 2008. Yawat:Yet Another Word Alignment Tool. In *Proceedings of the ACL-08: HLT Demo Session*, Columbus, Ohio.

T. Hirao, T. Fukusima, M. Okumura, C. Nobata, and H. Nanba. 2004. Corpus and evaluation measures for multiple document summarization with multiple sources. In *Proceedings of the 20th international conference on Computational Linguistics*, Geneva, Switzerland.
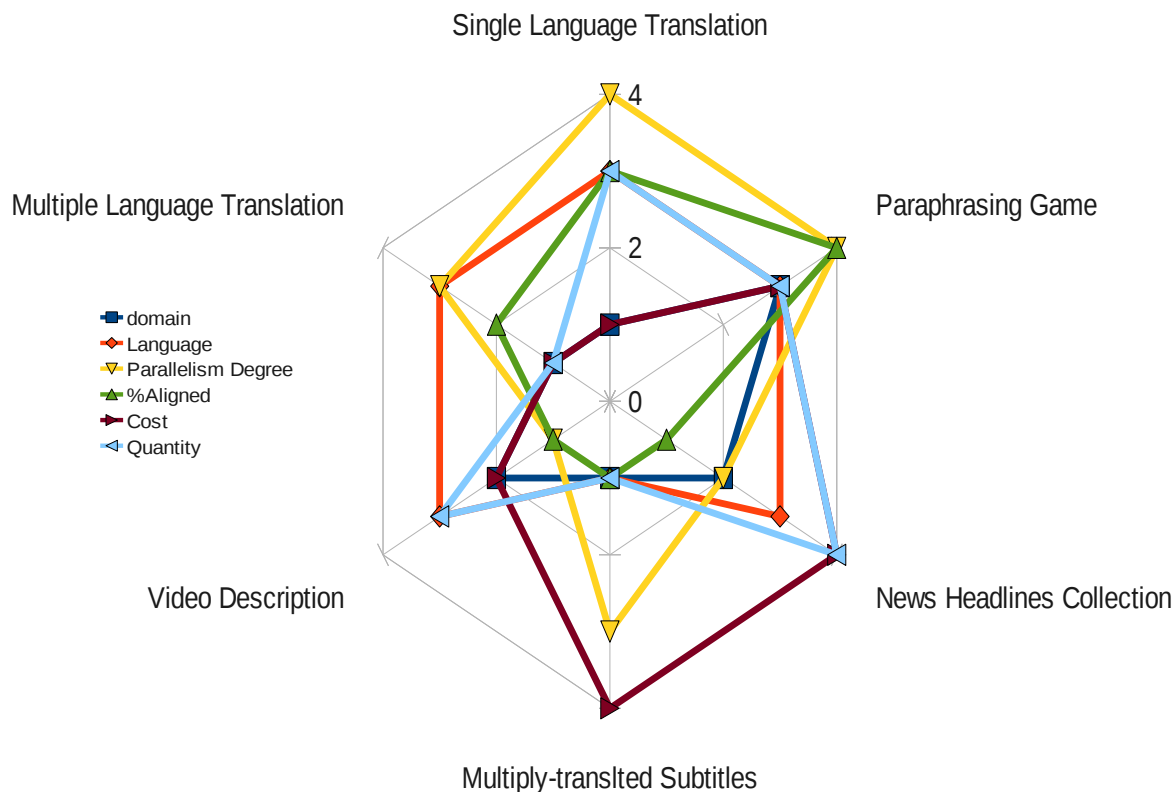
Figure 1: Various properties for each paraphrase acquisition approach

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods . *Computational Linguistics*, 36(3).

N. Madnani, P. Resnik, B.J. Dorr, and R. Schwartz. 2008. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of AMTA*, Waikiki, Hawai'i.

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In *Proceedings of LREC*, Valetta, Malta.

Eiichiro Sumita Mitsuo Shimohata and Yuji Matsumoto. 2004. Building a paraphrase corpus for speech translation. In *Proceedings of LREC*, Lisbon, Portugal.

Preslav Nakov. 2008. Improved statistical machine translation using monolingual paraphrases. In *Proceedings of ECAI*, Patras, Greece.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, Philadelphia, USA.

Marius Pasça. 2005. Mining paraphrases from self-anchored web sentence fragments. In *Knowledge Discovery in Databases: PKDD 2005*, volume 3721.

Chris Quirk, Chris Brockett, and William B. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP*, Barcelona, Spain.

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word Lattices for Multi-Source Translation. In *Proceedings of EACL*, Athens, Greece.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, Boston, USA.

Jörg Tiedemann. 2007. Building a multilingual parallel subtitle corpus. In *CLIN17*, Leuven, Belgium.

Sander Wubben, Antal van den Bosch, Emiel Krahmer, and Erwin Marsi. 2009. Clustering and matching headlines for automatic paraphrase acquisition. In *EWNLG*, Athens, Greece.

Shiqi Zhao, Haifeng Wang, Xiang Lan, and Ting Liu. 2010. Leveraging Multiple MT Engines for Paraphrase Generation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China.