Chinese Whispers: Cooperative Paraphrase Acquisition

Matteo Negri[1], Yashar Mehdad[1,2], Alessandro Marchetti[3], Danilo Giampiccolo[3], Luisa Bentivogli[1]

FBK-irst[1], University of Trento[2], CELCT[3]
Povo-Trento, Italy
negri@fbk.eu, mehdad@fbk.eu, amarchetti@celct.it, giampiccolo@celct.it, bentivo@fbk.eu

Abstract

We present a framework for the acquisition of sentential paraphrases based on crowdsourcing. The proposed method maximizes the lexical divergence between an original sentence *s* and its valid paraphrases by running a sequence of paraphrasing jobs carried out by a crowd of non-expert workers. Instead of collecting direct paraphrases of *s*, at each step of the sequence workers manipulate semantically equivalent reformulations produced in the previous round. We applied this method to paraphrase English sentences extracted from Wikipedia. Our results show that, keeping at each round *n* the most promising paraphrases (*i.e.* the more lexically dissimilar from those acquired at round *n-1*), the monotonic increase of divergence allows to collect good-quality paraphrases in a cost-effective manner.

Keywords: Crowdsourcing, Paraphrasing, Resources

1. Introduction

Paraphrase acquisition has received a great deal of attention in recent years. This is due to the number of NLP applications where different formulations of the same meaning are potentially useful. Among others, these include question answering (Hermjakob et al., 2002; Ravichandran and Hovy., 2002; Negri et al., 2008), textual entailment recognition (Hickl et al., 2006; Mehdad et al., 2011), information extraction (Banko and Etzioni, 2008), statistical machine translation (Callison-Burch et al., 2006), and machine translation evaluation (Kauchak and Barzilay, 2006).

Moving from early approaches based on costly manual work done by expert annotators, a variety of automatic acquisition methods has been proposed. Such methods alternatively focused on the exploitation of: *i)* mono/bi-lingual corpora, either parallel or comparable (Barzilay and McKeown, 2001; Bannard and Callison-Burch, 2005), *ii)* single monolingual corpora (Lin and Pantel, 2001; Szpektor et al., 2004; Bhagat and Ravichandran, 2008), or *iii)* the redundancy of the Web (Ravichandran and Hovy., 2002).

More recently (Chen and Dolan, 2011) proposed an acquisition methodology based on crowdsourcing, and defined a new evaluation metric (PINC - Paraphrase In N-gram Changes) to measure lexical divergence between source sentences and paraphrases. Two assumptions underlying (Chen and Dolan, 2011) are that: *i)* crowdsourcing is a viable approach to paraphrase acquisition, but *ii)* directly asking workers to paraphrase texts is not promising, since the task would be biased by the lexical or word order choices of the source sentences. To overcome this problem, they collected one-sentence descriptions of actions occurring in short video clips, and used PINC to measure their lexical dissimilarity. Then, in order to verify the usefulness of the resulting paraphrase corpus, they used the collected material to build a paraphrase system by training English to English translation models using Moses. Semantic adequacy and lexical dissimilarity were respectively measured with BLEU and PINC.

Despite the good results reported, two limitations might reduce the effectiveness of this method, namely: *i)* the machinery (*e.g.* thousands of video segments) needed to setup the acquisition, and *ii)* the fact that the collected material is parallel, but not necessarily semantically equivalent (*e.g.* "*A man dredges meat in bread crumbs*" is not a real paraphrase of "*A woman is adding flour to meat"*). Starting from the first assumption of (Chen and Dolan, 2011) about the viability of crowdsourcing for paraphrase acquisition, the main contribution of our work is to show that the issues motivating the second assumption (i.e., that directly asking workers to paraphrase texts is not promising) can be easily bypassed overcoming the aforementioned limitations. In particular, we describe a cheap and fast method for crowdsourcing paraphrase acquisition that:

1. Avoids the burden of setting up complex acquisition procedures (*e.g.* involving jobs like video-captioning);
2. Presents workers with a real paraphrasing task, but minimizes the impact of the lexical bias due to source sentences' wording;
3. Results in the acquisition of fully semantically equivalent paraphrases of long sentences, featuring large lexical divergence.

The material collected with our methodology (636 paraphrases of 100 original English sentences extracted from Wikipedia) is freely available for research purposes at: `http://www.celct.it/resourcesList.php`. Furthermore, the collected paraphrases have been used as a basis to create an English corpus of multi-directional entailment pairs and, after translation in several languages, a cross-lingual textual entailment corpus (Mehdad et al., 2010), which is being used in the Cross-lingual Textual Entailment for Content Synchronization Task organized within the SemEval-2012 evaluation campaign (Task#8[1]).

[1] http://www.cs.york.ac.uk/semeval-2012/task8/

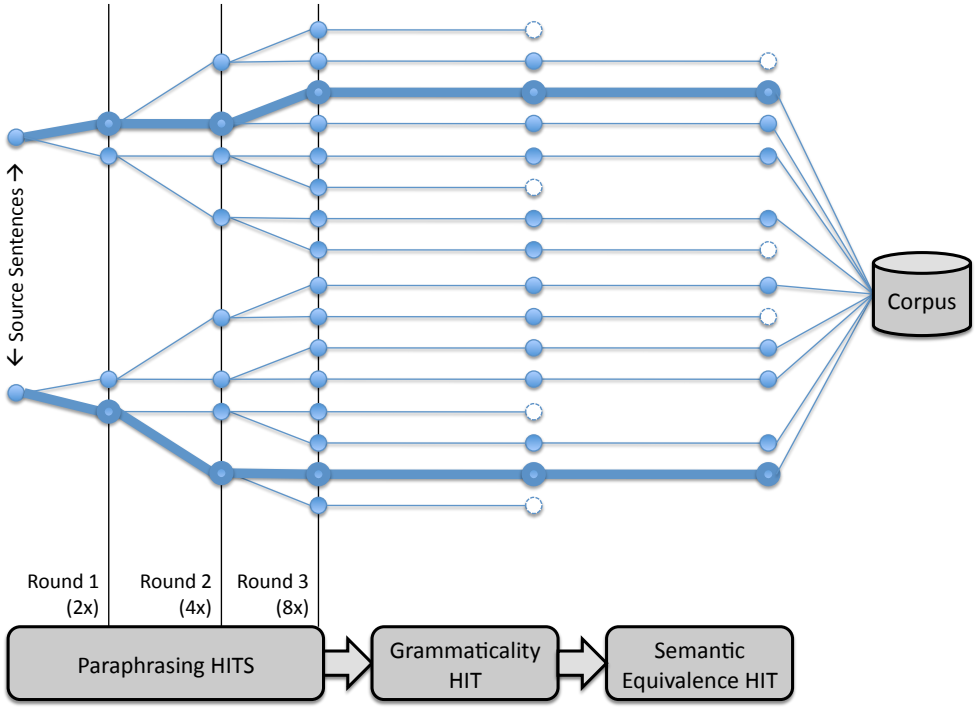

Figure 1: Data collection pipeline. After three rounds of paraphrasing, sentences that are both *grammatically correct* and *semantically equivalent* to the source texts are retained and stored in the paraphrase corpus. Thick lines represent paraphrases with highest lexical divergence according to a given metric.

2. Data collection pipeline

The paraphrase acquisition procedure started from 100 sentences extracted from randomly selected Wikipedia and Wikinews[2] articles. To reach such number, a larger amount of candidate sentences have been filtered to retain only those meeting two requirements:

- *Length*. Assuming that for short sentences the number of possible paraphrases is limited, from the selected articles we automatically extracted only sentences of at least 15 words;
- *Self-containment*. Assuming that in order to be easily modified into valid paraphrases a text has to be fully understandable, only sentences without external references have been retained. To this aim only the first sentence of each article has been considered, retaining those that do not contain anaphoric expressions.

The collected sentences were used to create different types of Human Intelligence Tasks (HITs) routed to Amazon Mechanical Turk[3] workforce through the CrowdFlower[4] interface. In order to keep the task feasible and maximize quality control over the collected data we adopted the "divide and conquer" approach described in (Negri et al., 2011; Negri and Mehdad, 2010). Under this framework, we decomposed the paraphrase generation task in a pipeline of simple subtasks that are easy to explain and execute, and suitable for the integration of a variety of runtime control mechanisms (regional qualifications, gold units, "validation HITs"). Our paraphrase acquisition pipeline (see Figure 1) contains three types of HITS. The first type of HIT ("Paraphrasing", depicted in Figure 2) aims at collecting semantically equivalent variants of a given sentence. As shown in Figure 2, a quality control mechanism (beside the regional qualifications common to all our HITs), paraphrasing HITs present workers with validation (*i.e.* YES/NO) questions about the semantic equivalence of two given sentences. Such *gold units* (*i.e.* sentences for which the correct judgement is known) represent a powerful mechanism to collect more accurate paraphrases by filtering out those obtained from workers that missed more than 30% of the gold questions. The second type of HIT ("Grammaticality", depicted in Figure 3) represents a quality check for the sentences collected from the paraphrasing task, and aims at filtering out the paraphrases that are not grammatically correct. As a quality control mechanism, the grammaticality job includes hidden gold units among the paraphrases sent to each worker. This mechanism allows to automatically filter out untrusted annotators (i.e., those that missed more than 30% of the gold answers). The third type of HIT ("Semantic Equivalence", depicted in Figure 4) asks to *"decide whether two English sentences contain the same information"*. This HIT, at the end of the pipeline, aims at filtering out grammatically correct sentences that are not paraphrases of the source sentence.Similar to the other jobs, also this HIT includes hidden gold units as a quality control mechanism.

[2]http://www.wikinews.org/
[3]https://www.mturk.com/
[4]http://crowdflower.com/

In this task you are asked to:
1) Answer a YES/ NO question about whether two English texts contain the same information.
2) **Change** the given text in order to obtain a new well-formed text containing the **same information**.
You can:
use synonyms (refusal/denial), different formats (3/three, US/United States), different expressions (Lebanese Minister/Minister of Lebanon), different structures (John ate the apple/The apple was eaten by John; John said that/according to John,) etc.

Do the following texts contain the same information?
• *Six people, including four U.N. staff working for the anti-narcotics department in Bolivia, have been killed in a plane crash in a remote area in the west of the country.*
• *Six people, also including four U.N. personnel working for the anti-narcotics division in Bolivia, have died in a plane crash in a isolated area in the west of the country.*
☐ YES ☐ NO

Copy the following text in the box below, and then substitute part of it preserving the meaning.
• *Two Katyusha Rockets launched from Taibeh, Lebanon, fell inside Kiryat Shmona, Northern Israel.*

...

Figure 2: Paraphrasing HIT.

In this task you are asked to decide if the given English sentence is **correct** or **incorrect**.
A sentence is to be considered correct if it does not contain grammatical mistakes, is well-formed and makes sense, even if it contains spelling or punctuation mistakes.
A sentence is to be considered incorrect if it contains grammatical mistakes, is not well-formed, or does not make sense (in a normal context).

• *Two Katyusha Rockets fired from Taibeh Lebanon, have struck inside northern Israel in the town of Kiryat Shmona.*

The sentence above is:
☐ Correct ☐ Incorrect

Figure 3: Grammaticality HIT.

In order to maximize lexical divergence between an original sentence *s* and valid, semantically equivalent reformulations {*p1,p2,...,pn*}, paraphrasing HITs are carried out through a sequence of rounds. At each round *n*, instead of generating direct paraphrases of *s*, workers manipulate the paraphrases produced at the previous round (*n-1*), in a sequence of cumulative modifications that resembles the children's "Chinese Whispers" game. In our experiment, we carried out three paraphrasing rounds and, at each round *n*, two different paraphrases of each input sentence were required, leading to 8 paraphrases for each of the 100 original sentences after the third round.

In this task you are asked to decide whether two English sentences contain exactly the **same information**.
Note that sentences may present different wording, but carry exactly the same information.
On the contrary, two sentences may have similar wording, but carry different information.

• *Two Katyusha Rockets fired from Taibeh Lebanon, have landed inside northern Israel in the town of Kiryat Shmona.*
• *Two Katyusha Rockets launched from Taibeh, Lebanon, landed within Kiryat Shmona in Northern Israel.*

Do the sentences above contain exactly the same information?
☐ YES ☐ NO

Figure 4: Semantic equivalence HIT.

3. Measuring divergence

We measured the lexical divergence of the acquired English paraphrases using 3 metrics: Lesk (Lesk, 1986), BLEU (Papineni et al., 2001), and PINC[5].
BLEU is a widely used precision oriented algorithm for evaluating the quality of machine translation output in comparison with reference translations. This score is based on the number of n-grams appearing in the output that also appear in the reference, normalized by the number of n-grams in the output. The final BLEU score is the average over n-gram scores, with values of n that typically cover the range from 1 to 4.
Lesk is a score that originally was proposed for word sense disambiguation. This score is calculated as the sum of the squares of the length of n-gram matches, normalized by dividing by the product of the string lengths.
PINC measures how many n-grams differ between the two sentences. This score computes the percentage of n-grams that appear in both the compared strings. PINC score is similar to the Jaccard distance, except that it excludes n-grams that only appear in the source sentence and not in the candidate sentence (Chen and Dolan, 2011).

	Round 1	Round 2	Round 3
Lesk	0.11	0.28	0.37
BLEU	0.18	0.31	0.38
PINC	0.16	0.28	0.34

Table 1: Average lexical divergence (Lesk, BLEU, and PINC scores) between source sentences and paraphrases, after each round of exhaustive acquisition.

In a first experiment we calculated the average distance of all the paraphrases acquired after each round. The scores reported in Table 1 coherently show a significant monotonic growth of lexical divergence with all the metrics, confirming the intuition that cumulative modifications are a viable

[5]Since Lesk and BLEU are originally *similarity* metrics, to compute divergence we subtracted the corresponding scores from 1.

solution for paraphrase acquisition. Although promising, it's worth recalling that our results are calculated after only three rounds of paraphrasing. At this stage it's difficult to say to what extent, and at what cost this exhaustive acquisition can monotonically improve.

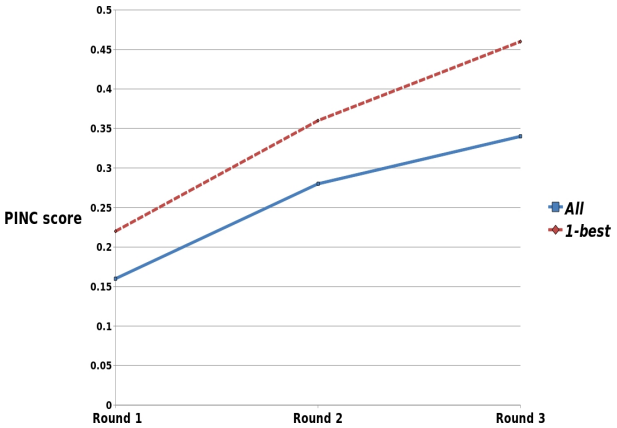

Figure 5: Lexical divergence (PINC score). Exhaustive (continuous line) and 1-best acquisition (dashed lines).

In a second experiment we used the PINC score to select, for each source sentence, the most lexically divergent paraphrase after each round, and send it to the following round. In principle, this solution has several advantages as it allows to: *i)* avoid the quadratic growth of exhaustive acquisition, *ii)* maximize lexical divergence by keeping the most promising paraphrases, *iii)* eventually saving time and money in favour of more paraphrasing rounds. As can be seen from Figure 5, this solution (dashed line) significantly increases lexical divergence compared to exhaustive acquisition (continuous line). Moreover, the corresponding curve shows a steep monotonic growth that suggests the possibility to further increase the divergence with few additional rounds. Although these scores are still below those reported in (Chen and Dolan, 2011), it's worth mentioning that:

- The "Grammaticality" and "Semantic Equivalence" HITs guarantee that what we collect are correct paraphrases of the source sentences;
- Such paraphrases are different in nature to the quasi-paraphrases collected through video captioning jobs. On one side, video captioning HITs allow to collect similar but often semantically different texts (*e.g. "A man dredges meat in bread crumbs* is is similar, but not equivalent to *"A woman is adding flour to meat*). On the other side, our pipeline is designed to retain sentences that maximize the semantic equivalence with the original texts;
- Their length and lexical/structural variability (see Example 6 in Table 2) has a great potential in a variety of applications.

4. The Resulting Paraphrase Corpus

After three rounds, 820 paraphrases of the 100 original sentences were collected, i.e. about 8 paraphrases for each original sentence.[6] Out of this total, the Grammaticality and Semantic Equivalence HITs respectively filtered out 117 (14%) and 67 (9.5%) sentences, showing an overall good quality of the collected paraphrases both in terms of syntax and meaning equivalence with the original texts. In the end, a corpus of 636 paraphrases was obtained.
The cost of running the whole pipeline was around $170, corresponding to 0.27$ for each final paraphrase[7].
As far as length is concerned, the original sentences ranged from a minimum of 18 words to a maximum of 59 words, with an average of 33.15 words. The final paraphrases where averagely 32.69 word long, meaning that the paraphrases were slightly, though not significantly shorter than the original sentences.

5. Paraphrase Analysis

A number of manual checks were carried out by expert annotators in order to verify the quality of the obtained paraphrase. More specifically, we checked:

(i) The kind of modifications made to the original text to obtain the paraphrases;

(ii) The grammaticality of the final output and semantic equivalence with the original sentences;

(iii) The effectiveness of keeping at each acquisition round the most promising paraphrases, i.e. those with the highest PINC score.

Table 2 lists some of the manually checked samples; hereafter the examples are quoted by referring to their number in this table.

5.1. Type of modifications

In order to analyze the kind of modifications performed during the paraphrase process, a sample of paraphrases was checked throughout the entire three-round process, starting from 30 final paraphrases and going back to the original sentences through the previous paraphrase rounds.
The analysis showed that the most common modification consisted in lexical substitution, often performed as a sequential replacement of single words in each round, as Example shows.
Additionally, in some cases also syntactic modifications were performed, such as negation, appositions, nominalizations/verbalizations, resulting in more complex paraphrase structures. For instance, in Example 2:

(i) First a synonym for *people* was used in Round 1;

(ii) Then the relative clause was substituted by an apposition through verbal nominalization in Round 2 (*was generated* was replaced by *consequence*);

[6] In the end 820, instead of 800 paraphrases were obtained due to the fact that crowdsourced jobs sometimes return a number of output items that is slightly larger than required, depending on the labour distribution mechanism internal to MTurk.

[7] Considering also the paraphrases created in the first and second round, a total of 14 paraphrases for each original sentence is obtained, which means that the actual cost per paraphrase is lower, about 0.12$.

#	SOURCE	PARAPHRASE	QUALITY
1	*During the Second World War, Agatha Christie wrote two novels, Curtain and Sleeping Murder, intended as the last cases of these two great detectives, Hercule Poirot and Jane Marple, respectively.*	*During the Second World War, Agatha Christie **authored [Round 3]** two novels, Curtain and Sleeping Murder, intended as the **final [Round 1]** cases of these two great **investigators [Round 2]**, Hercule Poirot and Jane Marple, respectively.*	GOOD
2	*A tsunami that was generated in the South Pacific by a powerful undersea earthquake has killed at least 110 people.*	*A tsunami in the South Pacific, **result [Round 3] of a powerful undersea earthquake [Round 2]**, has killed at least 110 **persons [Round 1]**.*	GOOD
3	*In the face of demand for higher fuel efficiency and falling sales of minivans, Ford moved to introduce a range of new vehicles, including "Crossover SUVs" built on unibody car platforms, rather than more body-on-frame chassis.*	*Ford's introduction of a new range of vehicles (like "Crossover SUVs") that were built on unibody car platforms instead of the body-on-frame chassis, was in response to both plunging minivans sales and demands for greater fuel efficiency.*	GOOD
4	*The Gates of Alexander was a legendary barrier supposedly built by Alexander the Great in the Caucasus to keep the uncivilized barbarians of the north (typically associated with Gog and Magog) from invading the land to the south.*	*To prevent the uncivilized barbarians from the north (who were typically associated with Gog and Magog) from overrunning the land to the south, Alexander the Great is thought to have built, in the Caucasus, the legendary barrier referred to as the Gates of Alexander.*	GOOD
5	*Following the California gold rush that began in 1849 and the Australian gold rush that began in 1851, a larger amount of gold was put into commerce than could be easily absorbed by the normal channels.*	*More gold than could be absorbed with ease by normal paths was put into the economy following the gold rushes that began in California in 1849 and in Australia in 1851.*	GOOD
6	*It is possible to aim a spacecraft from Earth so that it will loop around the Moon and return to Earth without actually entering lunar orbit, following the so-called free return trajectory.*	*Free return trajectory of spacecraft aimed from the earth so that they circle the Moon but never enter its orbit is a possibility.*	GOOD
7	*China is the world's second largest economy, having grown robustly in the last two years driven by a 2009 governmental stimulus package of $586 billion and low interest rates offered by state-owned banks.*	*China is the world's second largest economy, having grown tremendously in the last two years driven by a 2009 government stimulus plan of $586 billion and low interest rates at state-owned financial institutions.*	BORDER-LINE
8	*The trial of eight retired police officers and two witnesses involved in a high-profile Welsh miscarriage of justice began in Swansea, south Wales.*	*A high-profile Welsh miscarriage of justice started in Swansea, south Wales with the trial of 8 ex-police officers and 2 witnesses.*	REJECTED
9	*Wikileaks, so far, have released under 300 of the quarter million plus diplomatic communications posted to them on a memory stick.*	*Up to now Wikileaks have released less than 300 of the quarter million and diplomatic communications posted to them on a memory stick.*	REJECTED
10	*Also known as Dot or Dottie, Parker was born Dorothy Rothschild to Jacob Henry and Eliza Annie Rothschild (née Marston) at 732 Ocean Avenue in the West End village of Long Branch, New Jersey, where her parents had a summer beach cottage.*	*Also called Dot or Dottie, Parker was born Dorothy Rothschild to Jacob Henry and Eliza Annie Rothschild (née Marston) at 732 Ocean Avenue in the West End village of Long Branch, New Jersey, where her parents were the owners of a cottage.*	REJECTED
11	*Also known as Dot or Dottie, Parker was born Dorothy Rothschild to Jacob Henry and Eliza Annie Rothschild (née Marston) at 732 Ocean Avenue in the West End village of Long Branch, New Jersey, where her parents had a summer beach cottage.*	*Also known as Dot or Dottie, Parker was born Dorothy Rothschild to Jacob Henry and Eliza Annie Rothschild (née Marston) at 732 Ocean Avenue, the West End village of Long Branch NJ where her parents had a cottage.*	REJECTED
12	*A gunman, identified by Norwegian media as Anders Behring Breivik, has shot and killed at least 85 people at a youth camp on the island of Utya in Norway.*	*A perpetrator, identified as Anders Behring Breivik, has shot and killed at least eightyfive people at a youth camp on the island of Utya in Norway.*	REJECTED
13	*A gunman, identified by Norwegian media as Anders Behring Breivik, has shot and killed at least 85 people at a youth camp on the island of Utya in Norway.*	*At a youth camp located on the island of Utoya, Norway, 85 people have been shot and killed by a gunman, identified as Anders Behring Breivik.*	REJECTED
14	*During the Second World War, Agatha Christie wrote two novels, Curtain and Sleeping Murder, intended as the last cases of these two great detectives, Hercule Poirot and Jane Marple, respectively.*	*During the World War II, Agatha Christie authored two novels, Curtain and Sleeping Murder, intended as the last cases of Hercule Poirot and Jane Marple.*	REJECTED

Table 2: Examples of crowdsourced paraphrases and corresponding quality judgments by experts

(iii) Finally a synonym for *consequence* was introduced in Round 3.

Particularly good results are represented by parahprases that combine a variety of modifications both at lexical and syntactic level, as Examples from 3 to 5 show; the best outcome was obtained when lexical and syntactical variation was performed while introducing also a significant difference in length between the source sentence and the final paraphrase, as in Example 6.

5.2. Grammaticality and semantic equivalence

A manual check on a sample of 100 paraphrases randomly selected from the corpus was carried out in order to further verify the actual grammaticality of the final output, and semantic equivalence with the original sentences. The results showed that 8 out 100 were not "perfect" paraphrases (examples from 7 to 14) for different reasons, namely:

(i) An improper synonym was used partially changing the meaning of the sentence (one borderline case, Example 7, where human annotators did not agree on whether "financial institutions" is an acceptable subtitution for "banks");

(ii) The modifications in the series of paraphrases changed the meaning of the initial sentence (one case, Examples 8);

(iii) The paraphrase was not a well-formed English sentence (one case, Examples 9);

(iv) Part of the information present in the original sentence was missing in the paraphrase (five cases, Examples from 10 to 14).

On the basis of such analysis, assuming that an additional 8% of sentences is to be discarded, we can conclude that after three rounds of paraphrases an average of almost six good paraphrases for each original sentence can be obtained with the proposed methodology.

5.3. PINC score and quality

An additional manual analysis was performed in order to check the effectiveness of keeping at each acquisition round only the candidate best paraphrases, i.e. those with the highest PINC score. For each of the 100 original sentences we selected the best paraphrase and another paraphrase with a lower PINC score. For each of the resulting 100 triples (original sentence/best-scoring paraphrase /lower-scoring paraphrase) experts annotators were asked which paraphrase was best. In 88% of the cases the highest PINC score corresponded to a *best* judgment by human assessors. In the remaining 12 cases, the best PINC scoring paraphrases were judged as worse. This was due to the fact that they were considered as not acceptable *per se*, basically for the same reasons found in the previous manual check on grammaticality and semantic equivalence (see 5.2.):

(i) The modifications in the series of paraphrases changed the meaning of the original sentence (2 cases);

(ii) The paraphrase was not a well-formed English sentence (1 case);

(iii) Part of the information present in the original sentence was missing in the paraphrase (9 cases).

These results basically correspond to the proportion of faulty paraphrases found in the previous manual check, except that the number of paraphares with missing information is higher (9 out of 100 vs 5). Anyway, this difference was expected, as high PINC scores mean high divergence from the original, and the probability that this distance is due to a partial mismatch of information between source sentence and paraphrase gets higher as the PINC score increases.

6. Conclusion

Despite the great potential of crowdsourcing has been demonstrated in a number of data acquisition/labelling tasks, sometimes its application for the direct acquisition of paraphrases has been considered a problematic issue. The difficulty of the task lies in the fact that directly asking workers to paraphrase texts would produce results that are biased towards the lexical or word order choices of the source sentences. To overcome this problem we proposed a crowdsourcing method that, although asking for direct paraphrases, maximizes lexical divergence between original sentences and valid (*i.e.* syntactically correct and semantically equivalent) reformulations. The acquisition is carried out through a sequence of rounds. At each round, workers manipulate the paraphrases produced in the previous round by other workers, in a sequence of cumulative modifications that resembles the childrens "Chinese Whispers" game.

The lexical divergence between the original sentences and the collected paraphrases was measured using different metrics (Lesk, BLEU, PINC) and the results obtained demonstrate the effectiveness of the method proposed. In particular, keeping at each round *n* the most promising paraphrases (*i.e.* the more lexically dissimilar from those acquired at round *n-1*), the monotonic increase of divergence allows to collect good-quality paraphrases in a cost-effective manner.

The paraphrases generated with the proposed methodology present several advantages and can be used in a variety of NLP scenarios. For example, since high dissimilarity is hard to handle by Textual Entailment systems, a corpus made up of this kind of original-paraphrase pairs represents a potentially useful resource both for training and testing. Furthermore such a set of paraphrases could be used in developing and evaluating systems which deal with other semantic tasks.

Acknowledgments

This work has been partially supported by the EC- funded project CoSyne (FP7-ICT-4-24853).

7. References

M. Banko and O. Etzioni. 2008. The tradeoffs between traditional and open relation extraction. In *Proceedings of ACL 2008*.

C. Bannard and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL 2005*.

R. Barzilay and K.R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL 2001*.

R. Bhagat and D. Ravichandran. 2008. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of ACL 2008*.

C. Callison-Burch, P. Koehn, and M. Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL 2006*.

D.L. Chen and W.B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of HLT-ACL 2011*.

U. Hermjakob, A. Echihabi, and D. Marcu. 2002. Natural language based reformulation resource and web exploitation for question answering. In *Proceedings of TREC-2002*.

A. Hickl, J. Williams, J. Bensley, K. Roberts, B. Rink, and Y. Shi. 2006. Recognizing textual entailment with lccs groundhog system. In *Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*.

D. Kauchak and R. Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of HLT-NAACL 2006*.

M.E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC '86*.

D. Lin and P. Pantel. 2001. Discovery of inference rules from text. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*.

Y. Mehdad, M. Negri, and M. Federico. 2010. Towards cross-lingual textual entailment. In *Proceedings of the NAACL HLT 2010*.

Y. Mehdad, M. Negri, and M. Federico. 2011. Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of HLT-ACL 2011*.

M. Negri and Y. Mehdad. 2010. Creating a bi-lingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazons Mechanical Turk*.

M. Negri, M. Kouylekov, and B. Magnini. 2008. Detecting expected answer relations through textual entailment. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*.

M. Negri, L. Bentivogli, Y. Mehdad, D. Giampiccolo, and A. Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of EMNLP 2011*.

K. Papineni, S. Roukos, T. Ward, and Zhu WJ. 2001. Bleu: a method for automatic evaluation of machine translation. In *Technical Report RC22176, IBM, 2001*.

D. Ravichandran and E.H. Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of ACL 2002*.

I Szpektor, H. Tanev, I. Dagan, and B. Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of EMNLP 2004*.