

Collecting Paraphrase Corpora from Volunteer Contributors

Timothy Chklovski

Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292 USA
timc@isi.edu

ABSTRACT

Extensive and deep paraphrase corpora are important for a variety of natural language processing and user interaction tasks. In this paper, we present an approach which i) *collects multiple paraphrases per given item from volunteers* and ii) *incentivises responsible contributions by volunteer contributors*. Our approach is to solicit paraphrases from Web volunteers, both collecting new paraphrases with no prompting and *asking contributors to guess partially obfuscated paraphrases*. To test the approach, we have implemented an online game, *1001 Paraphrases* (<http://ai-games.org/paraphrase.html>), and deployed it to collect 20,944 entries focused on paraphrases of 400 statements. The approach complements existing text extraction methods and has some inherent unique advantages. We present and motivate our design as well as share preliminary observations and lessons learned about the performance of the approach.

Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning – *knowledge acquisition*; I.2.7 Natural Language Processing

General Terms: Design, Algorithms

Keywords: Volunteer contributor-based knowledge acquisition, interfaces for knowledge elicitation, paraphrase corpora

INTRODUCTION

Natural language permits us to say nearly the same thing in a great many ways. This variability of the surface form without significant impact on the meaning can present difficulties for a variety of intelligent user interfaces and natural language applications which rely on interpreting user input to respond to requests, take the requested action, or to identify redundant or similar information in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

K-CAP'05, October 2-5, 2005, Banff, Canada.

Copyright 2005 ACM 1-59593-163-5/05/0010...\$5.00.

summarization applications. Paraphrasing knowledge allows canonicalization of natural language, and thus can set the stage for such tasks as interpreting a user request in user dialogues (Allen et al, 2001), summarization (Knight & Marcu 2000), as well as in the performance application which we have focused on, which facilitates matching of user input in speech-to-speech translation of a small number of possible statements in a specific domain.

This paper's contribution is twofold. First, we put forth and empirically investigate a novel approach to constructing paraphrase corpora. The approach both complements existing approaches and has some unique advantages. Second, this paper makes a contribution in the emerging area of collecting knowledge from volunteers (Chklovski 2003a, 2003b; Chklovski & Gil 2005; Chklovski, 2005; Chklovski and Mihalcea, 2002; Mihalcea and Chklovski, 2004). Namely, this paper proposes an approach to the challenge of *designing the acquisition activity to incentivise responsible contributions*.

The approach taken in the system we present, *1001 Paraphrases*¹, is to ask contributors to enter new paraphrases and collect additional paraphrases by *asking contributors to guess partially obfuscated already known paraphrases*. By design, the approach both attracts contribution of novel paraphrases and keeps contributors focused on formulating and contributing appropriate paraphrases.

First, we briefly overview current leading approaches to collecting paraphrases, which are based on extraction from text corpora; we describe some unique advantages offered by the complimentary approach of collecting paraphrase knowledge from volunteer contributors over the Web. Next, we discuss how our deployed approach helps incentivise responsible contributions. After that, we present *1001 Paraphrases* which implements the approach, and discuss some preliminary observations and lessons learned from using *1001 Paraphrases* to collect 20,944 statements from approximately 1,300 visitors to the site,

¹ *1001 Paraphrases* is available online at <http://ai-games.org/paraphrase.html>

both in terms of success in constructing a corpus of paraphrases and incentivising responsible contributions.

COLLECTING PARAPHRASE CORPORA

The current leading approaches to collecting paraphrases are based on extracting paraphrases from text corpora. The extraction is typically done by identifying and aligning multiple versions of the same text in one language (e.g. Dolan et al., 2004), or multiple translations of the same text (e.g., Barzilay and Lee, 2003). Given a large text corpus, text extraction can potentially extract a large volume of paraphrase pairs. However, typically only few paraphrases for any given expression are extracted.

Turning to contributors, on the other hand, can allow collection of very many variants for a given expression, by soliciting much input about the existing items. Additionally, volunteers can potentially provide paraphrases for items not readily available in multiple text corpora – for example, paraphrases of statements typical to spoken, informal dialogue, rather than the more abundant text such as news corpora. Volunteers, in generating plausible paraphrases, also provide some degree of confidence in each paraphrase. The confidence in a pair of expressions being paraphrases of each other can be increased by asking additional volunteers and verifying that the paraphrase is consistently generated by them. Finally, obtaining paraphrases by text extraction often relies on (sometimes indirect) alignment at the word level of the two texts which contain paraphrases. Such alignment-driven extraction tends to yield paraphrases which share many words with each other. Automatically extracting paraphrases which are very distinct from each other at the word level but are still very close in meaning has so far proven difficult (Bill Dolan, personal communication).

Of the above strengths of collection from volunteers, the ability to focus on paraphrases for a given item and the selectable degree of validation are useful in applications such as the limited-phrasebook machine translation, the application which we have so far targeted with our deployment of *1001 Paraphrases*. At the same time, the approaches may complement each other, with the text extraction providing several high-confidence seed items, and/or with volunteer-contributed paraphrases being used to bootstrap extraction from text corpora.

Specific Application of Paraphrase Collection

1001 Paraphrases has been deployed to collect training data for a machine translation system which needs to recognize paraphrase variants of specific target expressions (Narayanan et al., 2003). The initial objective of the translation system is to allow an English-speaking doctor in a foreign country to communicate a limited number of statements and questions to a non-English (Persian) speaking patient. The supported communication

in the targeted, research version of the system is limited to four hundred recognized statements, such as “do you have a fever?”, “how long have you had these symptoms?” and “this will help you”. The goal is to allow the doctor to say, in English, any paraphrase of any recognized statement into a hand-held device. The paraphrase then is to be matched to the correct statement, and the pre-stored translation of that statement is to be output in the output language. Due to the potential variability of the statements, a challenging stage is to map the statement made to the closest matching recognized statement. To assist with this step, it is helpful to have a large corpus of paraphrases of the allowed statements². *1001 Paraphrases* has been deployed to collect a large number of possible paraphrases for the four hundred recognized statements. Although the site has not been actively promoted, it has been visited by approximately 1,300 contributors (not all of whom actually engaged in contributing paraphrases), and collected 20,944 distinct paraphrases over 15 months. The recognized statements initially came with one or two paraphrases each, a total of 400 statements and 638 additional paraphrases. These statements and their paraphrases were used as the seeds.

INCENTIVISING RESPONSIBLE CONTRIBUTIONS WHEN COLLECTING FROM VOLUNTEERS

Collecting from masses of volunteers can potentially provide very large amount of corroborated, multi-perspective knowledge. Fully delivering on the potential of collection from volunteers depends on creating systems which can attract a large volume of contributions. To date, a number of systems have been fielded, including *Open Mind Common Sense (OMCS)*, (Singh et al. 2002), *Open Mind Indoor Common Sense (OMICS)*, (Gupta and Kochenderfer, 2004), *LEARNER* (Chklovski 2003a, 2003b), *LEARNER2* (Chklovski & Gil 2005; Chklovski, 2005), *Open Mind Word Expert (OMWE)*, (Mihalcea and Chklovski, 2004; Chklovski and Mihalcea, 2002), and the *ESP Game* (von Ahn & Dabbish, 2004).

To increase the contribution volume, most of the deployed systems have offered recognition or prizes to those who provide the most contributions. However, rewarding for volume alone can encourage irresponsible contributions by unscrupulous participants. Such participants may opt to provide lots of arbitrary and incorrect, but quick and easy to enter input. Furthermore, automatically rapidly and accurately assessing the quality of the previously unseen input is inherently difficult. The fundamental reason is that the collection systems are relying on contributors to expand the available knowledge; when a novel contribution comes in, it is very difficult to know that it is incorrect without potentially costly additional verification,

² For an approach to canonicalizing volunteer-contributed natural language statements, see (Chklovski, 2003).

even if it appears unusual in light of the knowledge already present.

An important motivation of need for incentivising responsible contributions is evidence of irresponsible contributions in the absence of such incentives. In deployed systems for collecting knowledge from volunteers which reward contributors solely based on the volume of data contributed, there have been instances of contributors “gaming” the system. In *Open Mind Word Expert (OMWE)*, a collection site on which contributors annotate instances of words with their senses (Mihalcea and Chklovski, 2004), a concerned contributor has contacted the organizers asking that statements contributed from his account be deleted. The contributor indicated that his little brother used his account to enter a lot of low-quality data to be rewarded for the high volume of contribution, disregarding the notice that entries by winning contributors will be spot-checked for quality. Another experimental web-browser-based knowledge collection activity presented pages of multiple choice questions. Some contributors quickly discovered that they could rapidly increase their scores by using the browser’s back button and re-submitting many times the same form filled out once, further garnering high “agreement” with themselves (Stork, 2003). Furthermore, the more engaging and game-like the knowledge collection interfaces become, the greater may be the temptation for contributors to cut corners to get ahead.

A large volume of invalid contributions can be detrimental to a collection effort. At the same time, quickly and reliably identifying invalid contributions is made challenging by the very nature of collecting *new* knowledge, because the collection system is constantly receiving statements which it did not already have and thus does not know whether to trust. Reliably validating complex, somewhat rare entries which are not likely to be often entered redundantly is also non-trivial. One approach to judging quality of statements is to wait until other contributors also enter it. While a statement having been entered several times may be indicative of its reliability, a statement which has been entered only once may still be acceptable and useful, just rarely contributed. Another approach is to use “validating contributors” who are presented with previously entered statements and express whether they agree or disagree with them. Yet, if a validating contributor disagrees with a statement, further investigation is needed to establish whether the original or validating contributor is to be trusted. Despite these issues, assessing quality of statements contributed by volunteers is important and should be used in conjunction with the approach we investigate here.

Rather than struggle with the contributors, we propose an approach which structures the collection activity so that contributors *want to* provide responsible contributions in the first place. To that end, we propose a simple approach

to incentivising high quality contributions. Our approach is to collect additional answers by *asking contributors to guess partially obfuscated seed or previously known answers*. The approach is applicable when a question has a number of valid answers, as in the case of paraphrases for a given statement or question. For each “question” or a prompting item (in our case, the expression to be paraphrased), the approach requires at least one valid answer to seed the collection. Such seed answers may be solicited from contributors separately or perhaps be the highest precision answers automatically extracted from text corpora. Our approach then allows collection of additional answers.

To allow collection of novel answers completely unrelated to the already collected answers, in our approach as we have deployed it, each user has to make an initial guess with the answers completely obfuscated. Since the contributor is playing the “game” of guessing the target answers, the contributor still has an incentive to make plausible guesses even when the answers are completely obfuscated. Additionally, by incorporating the elements of guessing and immediate feedback on success and failure we aimed to make the interaction more engaging.

The approach can also be used to collect entries other than paraphrases. For instance, it can be used to collect answers to questions about everyday objects. To collect such entries, a question such as “computers are used to _____” would be presented, collecting as answers such phrases as “compose email,” “send email,” “compose documents,” “view images” and so on. Previously established answers can be used as targets to be guessed. For situations in which answers are easier to guess, the awarded score may be tied to the number of target answers guessed correctly.

In addition to the above approaches of rewarding for volume of contributions with manual verification of quality, one piece of work in particular focuses on incentivising anonymous volunteer contributors to make responsible contributions. In their work on the *ESP game* von Ahn and Dabbish explore this issue in the context of acquiring text labels for images found on the Web (von Ahn & Dabbish, 2004). The *ESP game* randomly pairs contributors to simultaneously enter annotations of a selected image from the Web with the label being acquired when the two contributors enter the same label. Since the goal of the game is for two contributors to guess the same label, the game encourages responsible contributions. A significant difference between *1001 Paraphrases* and the *ESP game* is the *ESP game*’s emphasis on requiring agreement without additional hints. The *ESP game* has been designed to generate mostly single-word labels for images. Applying the approach to sentence-long paraphrases without additional hints may be difficult on sentence-long paraphrases, due to the variety of paraphrases. However, it may be interesting to explore an

approach which is a hybrid of our hint-based approach and of the approach taken in ESP game, with several contributors competing on being the first to correctly guess the paraphrase of a given item, or with contributors guessing each others' paraphrases by being given hints about them.

1001 PARAPHRASES

Although *1001 Paraphrases* is a general platform for collecting paraphrases, it has been deployed for collecting paraphrases for a specific research project. We introduce this specific project, and then describe the interface, interaction, and scoring in *1001 Paraphrases*.

Interface of *1001 Paraphrases*

The “game” consists of a contributor making multiple attempts at guessing any of the several partially obscured paraphrases for a displayed expression. Figure 1 shows a screen shot of the *1001 Paraphrases* interface. Displayed at the top is the expression to be paraphrased, in this case “*this can help you*”. The partially obscured target expressions to guess are shown in the hints box. The “...” in the hints indicates that one or more words have been obscured.

The contributor enters a paraphrase in the box titled “Another way to say it”. If the contributor correctly guesses one of the paraphrases, he is awarded the amount of points specified next to “you can win,” and the game proceeds to the next item for paraphrasing. Otherwise, the entered expression is added to the list of expressions already tried by this contributor, and a larger fraction of the words for the paraphrases to be guessed are revealed in the hints box. The number of points you can win is also decreased. The contributor may also choose to ask for a hint. Just as an unsuccessful guess, this reveals more words in the target expression, but decreases the number of points you can win. The hint functionality has been added to allow contributors to get more information when they cannot think of a guess. At the same time, number of points you can win is decreased to encourage guessing earlier, when little information available, increasing diversity of collected paraphrases and reflecting the difficulty of guessing. The contributor can also give up; the full answers are then revealed and a new item to paraphrase is chosen at random from the target set. No points are awarded or subtracted.

When a new item to paraphrase is presented, no hint words are provided at all for the first guess. This is done to allow collection of paraphrases which share no words with the known paraphrases.

Selection and Obfuscation of Hints

As mentioned earlier, our approach relies on presenting obfuscated answers. In the deployed version, how much and exactly how the answers are obfuscated is important,



Figure 1. A Screenshot of *1001 Paraphrases*

both for whether novel answers will be entered and for the user experience. In designing the obfuscation, we aimed to provide the setting to enter new spontaneous contributions and to prompt contributors' thinking rather than to constrain it. While the hints may potentially bias new contributions towards the ones already collected, as the example in Figure 1 illustrates, the partial hints such as “this could ...” and “... help” for paraphrases of “*this can help you*” still allow much variation in the input.

The mechanism we have selected is to show a certain percentage of the words in the expression, replacing the runs of omitted words with “...”. For the first guess, we do not reveal any text of the hint, calling for a complete guess from the contributor. For the second guess, 66% of all words (regardless of part of speech etc) are obfuscated. For third and subsequent guesses, 33% of all words are obfuscated. We do not indicate the actual number of words in the expression by design. If we did, contributors may not enter some paraphrases because they would not “fit,” and this self-censorship would run counter to the goal of the collection.

Entering Variants of a Paraphrase

In designing *1001 Paraphrases*, we anticipated that it may be frustrating to contributors to think of many variants of paraphrases but to be able to enter only one as their guess. Restricting the number of paraphrases a contributor can enter is also counterproductive to the goal of the collection.

To address this issue, *1001 Paraphrases* supports compact entry of many variants of a paraphrase at once. All the entered variants count as one guess; the contributor wins the round if any variant matches. The variants are entered by using special symbols, such as “/” (for alternation), “_” (for keeping entities together in alternations), and “(”, and “)” for enclosing optional parts in the paraphrases. For example, entering “hello/hi there” expands to two phrases:

“hello there” and “hi there.” In designing the language for entering multiple guesses at once, we aimed to balance simplicity and expressiveness.

Here is an example exercising all of the allowed constructs:

`how_are_you_doing/is_it_going (today/this_morning)?`

This example expands to the following 6 questions:

```
how are you doing?
how are you doing today?
how are you doing this morning?
how is it going?
how is it going today?
how is it going this morning?
```

The contributed multi-paraphrase expressions are not used as hints. Only instances of expansions are used as hints.

In practice, 3.8% of the contributed paraphrases used this functionality. We speculate that many contributors may have been unaware of this functionality, because one had to visit the “help” page to learn about it, and server logs indicate that few visitors did. Also, perhaps more contributors would use this functionality to gain an edge in their scores if the activity was made more engaging or competitive (for example, by introducing into the game competition against time and maintaining a high-scores list).

PARAPHRASES COLLECTED

Although the seed set contained 1,038 statements (400 distinct recognized expressions and 638 paraphrases for them), Contributors provided many more by playing the game. Of the 20,944 contributions, some were identical, because different contributors made the same guesses. In all, 14,850 distinct entries were collected. By way of illustration, Table 1 shows all the statements entered for the statement “this will help you.” The seed paraphrases were “this’ll help” and “this will be of help.” Of 22 new contributions, the majority were good although some, such as “try this” or “it’s healthy” were further off in meaning, and one, “nice going” seems not relevant.

Table 1. All 26 distinct statements collected for “this will help you.”

Paraphrase	# times contributed	Paraphrase	# times contributed
try this	6	its healthy	2
this should do the trick	5	this could be of help	1
this will help you	3	this will make it better	1
this will help	3	this could be better	1
this should help	3	this makes you feel better	1
this will do the trick	3	this should do	1
this'll help	2	this will fix the problem	1
this will be of help	2	this should fix the problem	1

this should make it better	2	good for you	1
this should be better	2	this may help	1
this can help you	2	that should do the trick	1
this should help you	2	nice going	1
this is better	2		

For longer statements, such as “we will begin the operation as soon as we can,” the majority of paraphrases were still accurate, but some paraphrases omitted some of the information. For example, paraphrases such as “we will be ready soon” or “will start as soon as possible.” Although such omissions may be undesirable in other circumstances, they were mostly acceptable for our target application.

The paraphrases collected from volunteers were also used as part of the training data for the speech-to-speech application. The training data also contained paraphrases from other sources (manually authored by groups of users). The more training data was used, the better the resultant system performed. The detailed evaluation of the paraphrase data collected from volunteers with extracted data and the impact on application performance are subject of future work.

Examining the paraphrases collected to date suggests that contributors indeed tended to make responsible contributions. Of the contributed 20,944 paraphrases, only 0.23% contained swear words, and 0.12% were nonsense entries. By contrast, in our other work on collecting world knowledge from volunteers which did not include mechanisms for incentivising responsible contributions, the number of “noise” entries containing swear words or nonsense entries such as random key presses hovered at around 10%³. The relative scarcity of noise entries in the Paraphrase game may be indicative of contributors being more incentivised to comply with the activity goals overall. A much larger fraction of the statements, approximately 5%, contained misspellings. The rate of misspellings in other systems we have deployed was roughly similar (Chklovski 2003a, Chklovski, 2005).

CONCLUSIONS

Extensive and deep paraphrase corpora are important for a variety of natural language processing and user interaction tasks. One key contribution of our work lies in proposing and empirically investigating an approach to collecting paraphrase corpora from volunteer contributors.

A second contribution of our work is a method for incentivising potentially unscrupulous volunteers to make responsible contributions. Rather than reward for pure volume, the approach rewards for correctly guessing obfuscated previously known answers. This approach can

³ See (Chklovski, 2005) for data on acceptability of the data and overlap with meronymy (part-of) data in WordNet.

be applied beyond collection of paraphrase knowledge to collecting semantic relations, and other types of contributions.

Preliminary analysis of deploying the approach to collect paraphrases indicates that a significant number of novel, useful paraphrases were indeed collected with the approach, and that there was little evidence of non-compliant contributions.

In future work, we intend to compare in detail the paraphrase collections collectable from volunteers with the collections extractable from text. We also intend to explore collecting paraphrases for sentence fragments, as well as extend the approach to acquiring linguistic knowledge more specific than paraphrases such as entailment and paraphrases not readily available from textual corpora.

ACKNOWLEDGMENTS

I thank Kevin Knight and Emil Ettelaie for providing seed topics and ongoing evaluation of the collected data in the speech-to-speech system. I thank Yolanda Gil for valuable discussions on this work and paper.

REFERENCES

Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *ACM CHI 2004*

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. "Towards Conversational Human-Computer Interaction," *AI Magazine* 22(4), pages 27-38, Winter, 2001

Barzilay R. and Lee, L. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proceedings of North American Chapter of the Association for Computational Linguistics and Human Language Technology, (NAACL-HLT)*, 2003.

Belasco, A., Curtis, J., Kahlert, R., Klein, C., Mayans, C., Reagan, P. 2002. Representing Knowledge Gaps Effectively. In *Practical Aspects of Knowledge Management, (PAKM)*, Vienna, Austria, December 2-3.

Chklovski, T. and Mihalcea, R. 2002. Building a Sense Tagged Corpus with Open Mind Word Expert. In *Proceedings of the Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, Association for Computational Linguistics (ACL) 2002. pp. 116-122

Chklovski, T. 2003a. *Using Analogy to Acquire Commonsense Knowledge from Human Contributors*, PhD thesis. MIT Artificial Intelligence Laboratory technical report AITR-2003-002

Chklovski, T. 2003b. LEARNER: A System for Acquiring Commonsense Knowledge by Analogy. In *Proceedings of Second International Conference on Knowledge Capture (K-CAP 2003)*.

Chklovski, T. and Gil, Y. 2005. An Analysis of Knowledge Collected from Volunteer Contributors. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*.

Chklovski, T. 2005. Designing Interfaces for Guided Collection of Knowledge about Everyday Objects from Volunteers. In *Proceedings of Conference on Intelligent User Interfaces (IUI-05)* San Diego, CA

Dolan, W. B., Quirk, C., and Brockett, C. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Gupta, R., and Kochenderfer, M. 2004. Common sense data acquisition for indoor mobile robots. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*.

Knight, K. and Marcu, D. 2000. Statistics-based summarization – Step one: Sentence compression. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-00)*.

Mihalcea, R., and Chklovski, T. 2004. Building Sense Tagged Corpora with Volunteer Contributions over the Web. In *Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing*, Nicolas Nicolov and Ruslan Mitkov (eds), John Benjamins Publishers.

Narayanan, S., Ananthkrishnan, S., Belvin, R., Ettelaie, E. Ganjavi, S. Georgiou, P., Hein, C., Kadambe, S., Knight, K., Marcu, D., Neely, H., Srinivasamurthy, N., Traum, D. and Wang, D. 2003. Transonics: A Speech to Speech System for English-Persian Interactions. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (IEEE ASRU)*.

Stork, D. 2003. Invited talk at the *Workshop on Distributed and Collaborative Knowledge Capture (DC-KCAP)*, held in conjunction with the *International conference on Knowledge Capture (K-CAP 2003)*.