

Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results

Workshop Programme

14:00 – 14:25 – Introduction by Workshop Chair

Christopher Cieri, *Novel Incentives in Language Resource Development*

14:25 – 15:15 – Novel Incentives in Data Collection and Requisite Processing

Nick Campbell, *Herme & Beyond; the Collection of Natural Speech Data*

Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, Tomoya Mizumoto, *FKC Corpus: a Japanese Corpus from New Opinion Survey Service*

15:15 – 16:05 – Novel Incentives and Workflows for Annotation

Kara Greenfield, Kelsey Chan, Joseph P. Campbell, *A Fun and Engaging Interface for Crowdsourcing Named Entities*

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz and Chris Madge, *Novel Incentives for Phrase Detectives*

16:05 – 16:30 – Afternoon Coffee Break

16:30 – 17:20 – Understanding and Exploiting Data from Alternative Sources

Na'im Tyson, Jonathan Roberts, Jeff Allen, Matt Lipson, *Evaluation of Anchor Texts for Automated Link Discovery in Semi-structured Web Documents*

Maxine Eskenazi, Sungjin Lee, Tiancheng Zhao, Ting Yao Hu, Alan W Black, *Unconventional Approaches to Gathering and Sharing Resources for Spoken Dialog Research*

17:20 – 18:00 – The Future of Incentives, Workforces, Workflows and Data Exploitation

Mark Liberman, *Oral Histories: Linguistic Documentation as Social Media*

Wrap-Up and Discussion

Table of Contents

Christopher Cieri, <i>Novel Incentives in Language Resource Development</i>	1
Nick Campbell, <i>Herme & Beyond; the Collection of Natural Speech Data</i>	7
Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, Tomoya Mizumoto, <i>FKC Corpus: a Japanese Corpus from New Opinion Survey Service</i>	11
Kara Greenfield, Kelsey Chan, Joseph P. Campbell, <i>A Fun and Engaging Interface for Crowdsourcing Named Entities</i>	19
Massimo Poesio, Jon Chamberlain, Udo Kruschwitz and Chris Madge, <i>Novel Incentives for Phrase Detectives</i>	23
Na'im Tyson, Jonathan Roberts, Jeff Allen, Matt Lipson, <i>Evaluation of Anchor Texts for Automated Link Discovery in Semi-structured Web Documents</i>	27
Maxine Eskenazi, Sungjin Lee, Tiancheng Zhao, Ting Yao Hu, Alan W Black, <i>Unconventional Approaches to Gathering and Sharing Resources for Spoken Dialog Research</i>	32
Mark Liberman, <i>Oral Histories: Linguistic Documentation as Social Media</i>	38

Author Index

Allen, Jeff	27
Black, Alan W.	32
Campbell, Joseph P.	19
Campbell, Nick	7
Chamberlain, Jon	23
Chan, Kelsey	19
Cieri, Christopher	1
Domoulin, Mathieu	11
Eskenazi, Maxine	32
Greenfield, Kara	19
Hu, Ting Yao	32
Kruschwitz, Udo	23
Lee, Sungjin	32
Liberman, Mark	38
Lipson, Matt	27
Madge, Chris	23
Mitsuzawa, Kensuke	11
Mizumoto, Tomoya	11
Nakashima, Masanori	11
Poesio, Massimo	23
Roberts, Jonathan	27
Tauchi, Maito	11
Tyson, Na'im	27
Zhao, Tiancheng	32

Preface/Introduction

Despite more than two decades of effort from many research groups and large data centers, the supply of LRs falls far short of need even in the languages with the greatest number of speakers, controlling the largest shares of the world economy. For languages with less international recognition, resources are scarce, fragmentary or absent. Recent programs such as DARPA LORELEI recognize and attempt to address this gap but even they will provide only core resources for a few dozen languages, a small proportion of the >7000 currently in use worldwide.

In Language Resource (LR) development the commonest incentives for contributors are monetary. Whether motivated by convenience or ethical beliefs, that bias limits the Human Language Technology (HLT) community's ability to collect data and understand how different incentives impact collection. Because linguistic innovation is effectively limitless, relying upon a limited resource, monetary compensation, to generate the data needed to document the world's language is certain to fall short. Instead LR developers and users must develop and employ incentives that scale beyond the budget of a 3- or 5-year program.

A few HLT projects have employed alternate incentives. *Phrase Detectives* provides entertainment, challenge and access to interesting reading in exchange for anaphora annotation. *Herme* gave participants the unusual experience of interacting verbally with a tiny, cute robot while recording their interactions. *Let's Go* mediates access to Pittsburgh Port Authority Transit bus schedules and route information while recording the interactions to improve system performance in real world situations especially for 'extreme' users such as non-native and elderly speakers.

However, outside our field, collections employ variable incentives to much greater effect, creating massive data resources. *LibriVox* offer contributors the chance to create audio recordings of classic works of literature, develop their skills as reader and voice actors, work within a community of similarly minded volunteers and enable access to the blind, illiterate and others. *Zooniverse* includes linguistic exercises such as the transcription of originally hand written bird watching journals and artists' diaries or of the typewritten labels of insect collections. Social media has employed a wide range of incentives including:

- access to information and entertainment
- possibilities for self-expression, sharing and publicizing intellectual or creative work
- chances to vent frustrations or convey thoughts, sometimes anonymously
- forums for socializing; exercises which develop competence that may lead to new prospects
- competition, status, prestige, and recognition
- payment or discounts in real and virtual worlds
- access to services and infrastructure based on contributions
- novel experiences and improved interactions, for example in a customer service encounter
- opportunities to contribute to a greater cause or good

While lagging behind in the use of novel incentives, HLT researchers have productively used crowdsourcing to lower collection and annotation costs and developed techniques for customizing tasking to meet the capacity of the crowd and fusing highly variable results into data sets that advance technology development. Similar techniques apply to the use of alternate incentives in collecting data from a non-traditional workforce.

This half-day workshop will open the discussion on incentives in data collection describing novel approaches and comparing with traditional monetary incentives. Related topics including: descriptions of projects that use the alternate incentives listed above or others; modifications of the data collection and annotation tasking or workflow to accommodate a new workforce, including crowdsourcing; techniques for exploiting the results of alternate incentives and novel workflows.

Novel Incentives in Language Resource Development

Christopher Cieri

Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
{ccieri} AT ldc.upenn.edu

Abstract

The gap between supply of and demand for Language Resources continues to impede progress in linguistic research and technology development, even in the face of immense international effort to create the requisite data and tools. This deficiency affects all languages in some way, even those with worldwide economic and political influence. Moreover, for most of the world's 7000 linguistic varieties the absence is acute. Current approaches cannot hope to meet the resource demand for even a reasonable subset of the languages currently spoken because they seek to document phenomena of great variability principally using resources, such as national funding, that are highly constrained in terms of amount, duration and scope. This paper describes efforts to augment the traditional incentives of monetary compensation with alternate incentives in order to elicit greater contributions of linguistic data, metadata and annotation. It also touches on the adjustments to workforces, workflows and post-processing needed to collect and exploit data elicited under novel incentives.

Keywords: novel incentives, workflows, language resources

1. Introduction & Motivation

Despite the immense contributions of worldwide data centers, national language corpus projects, government agencies, research groups and individual contributors, the supply of language resources still falls far short of demand. Human Language Technology (HLT) developers experience this shortfall not only in the average but also for every single human language. The METANET (2010) white paper series documents the language resources required to build the technologies needed to *future-proof* European language against *digital extinction*, that is to allow them to participate in an increasingly digital, information driven world. As the reports compare need to existing resources for EU languages, they demonstrate that no language, not even English, enjoys the full range and that “*21 out of 30 European languages could become extinct in the digital world*”. What is true for EU languages is at least as much true for the remainder of the world's languages.

Success in the digital domain is only one of many motivations for creating HLTs and the pre-requisite resources. A 2008 report from the International Association of Conference Interpreters warned: “*Ending a conflict and delivering emergency and humanitarian aid across language barriers represents a major challenge, for which few of the organisations entrusted with operations in the field are well equipped. This problem is compounded by the fact that there is a chronic shortage of interpreters in zones of crisis and war willing to work in the line of fire or in areas of natural disaster.*” Although technologies have the potential to streamline disaster relief, the delay between the onset of the disaster and the integration of the technology continues to thwart relief efforts: “*Effective communication in Haiti was confronted by language barriers and the limited utilization of technology. Media played an important role in*

communicating about the disaster relief effort to the international community, but their reporting at times included misinformation.” (Harvard Humanitarian Initiative 2011).

HLTs have a growing role – and will have a critical role in the future – in disaster relief. Varma et al. (2011) showed their potential by using natural language processing techniques to filter tweets, with 80% accuracy, according to whether they provided situational awareness. However, the system required training data to be annotated not only for situational awareness but also for subjectivity, formality, and personal versus impersonal viewpoint. In addition, their automatic processing included a part-of-speech tagger, which cannot be assumed to exist for most low resource languages. Indeed even the tokenizer, list of stop words, unigram and bigram frequencies are absent for many of the world's languages almost certainly some that will figure into future disasters.

A number of US government programs over the past several years have begun to address the need for HLTs and pre-requisite LRs to support disaster relief efforts. In 2011 the National Science Foundation (NSF) provided \$2.8M in support to the EPIC (Empowering the Public with Information in Crisis) project at U. Colorado and U.C. Irvine researching technologies to facilitate disaster relief communications. DARPA LORELEI is developing technologies to deal with disaster related communication in low resource languages. However such programs last for just a few years and provide their impressive array of resources for at most a few dozen languages. The 19th edition of the Ethnologue (Lewis, Simons, Fennig 2016) reports the tally of living languages to be 7,097 worldwide most of which lack the resources required by Varma et al.'s system and will not be the focus of LORELEI or any current program.

Finally, the societal need for multilingual technologies and enabling data extends well beyond commerce, defense, and disaster relief. A 2010 article published by the American Psychological Association echoed the growing need for greater translanguaging capability, which they characterized in terms of interpreters within counseling services.

In summary, current approaches to HLT and LR development will not meet the needs of human language technologies for the world’s languages or even an appreciable subset of them. In order to scale significantly beyond current production it will be necessary to revolutionize multiple aspects of LR development including the conceptualization of the tasking, the target workforces and their motivations, the workflows used to acquire data, metadata and judgments and the post-processing necessary to exploit next generation LRs in the development of HLTs.

2. An Incentives-Aware Model of Language Resource Creation

Each process that creates data or annotation used in linguistic research and technology development, whether it does so intentionally or in the service of some other goal, can be seen in terms of several interacting components: the task, the incentives offered, the workforce that the incentives attract, the workflow required to permit to workforce to complete the required task and the output. Different workforces are motivated by different incentives, require different tasking and workflows and produce different outcomes. All of these factors impact the researcher who would use the data as well as the organization that would collect it. Greenfield, Chan and Campbell (2016) provide an example: *“While annotators who have been trained as professional linguists are able to annotate accurately and consistently from dense annotation guidelines, the amateur annotators who serve as workers on crowdsourcing platforms are not similarly motivated to create the best annotations possible.”*

The Human Language Technology (HLT) communities are already familiar with found data types such as newswire and broadcast news that are created for purposes unrelated to HLT and rely upon workforces and workflows outside their control. For data types created specifically to support HLT research and development, common incentives include monetary compensation and in smaller scale efforts the potential to use the data for ones own research. However the conscious engineering of incentives, workforces and workflows to optimize output for a specific task is rather limited within the HLT LR production. There are obvious counter-examples. Much of the recent work on crowd-sourcing discusses the impact of factors such as HIT size and complexity, payment rate, and instructions on the quality of the outcome and design sophisticated interfaces to harness the wisdom of the crowd, and reduce cheating. However this valuable research relies principally on the incentive of monetary

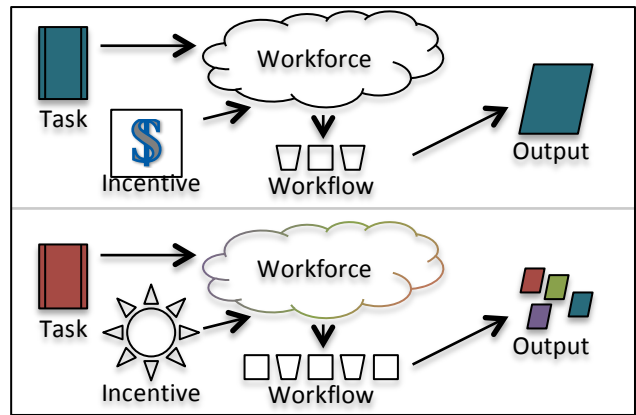


Figure 1: Different incentives attract different workforces that require different tasking and workflows and produce different outputs.

compensation. In much older work multiple LDC studies (Cieri et al. 2006, 2007) have reported on the relative effects of graduated pay scales, completion bonuses and random prizes upon performance in telephone collections. However, again, the incentives were principally pecuniary. In the next section we will review some very recent work within HLT communities in engineering incentives and/or engineering workflows to deal with data created under non-traditional incentives. These include cases of HLT development for industry where the specific combination of workforce, incentive and workflow is the target environment for the technology as well as other cases in which the environment has been engineered to provide data for some other purpose.

3. Incentives in Language Resource Development for HLT

In the sections below we focus on several very recent efforts with the HT communities to make use of novel incentives in data collection and annotation including new workflows and post-processing necessary to use such data in system building.

3.1. Collection

Campbell (2016) reports on a number of data collections intended principally to support the development of systems capable of producing expressive speech. These collection efforts experimented with a variety of incentives and adjusted to the different characteristics of the output. In addition to any monetary compensation, additional motivations included access to the resulting data for research purposes, sustenance, curiosity, fun, ability to keep the recording device used and opportunities for unusual social interactions including apparent conversations with a robot and extended interactions with colleagues outside the lab. The data resulting from these studies naturally varied along many dimensions including the proportions of regional speech, emotion, non-speech vocalizations and contact events. Based on his own experience in both worlds, Campbell also emphasizes the growing divide between academic and industrial HLT research especially in terms of data

volumes. From our perspective in this paper, the motivations of acquiring some product or service can be seen as leading commercial customers to provide vast quantities of ‘data’ to HLT researchers in industry.

Continuing that theme, Mitsuzawa et al. (2016) describe their efforts to process product and company reviews from the Fuman Kaitori Center. Like many developers in industry they enjoy a reduced train-test mismatch because the data they use to build their systems is quite similar to, or an earlier instantiation of, the data their system will ultimately process. Consumers post their reviews initially to communicate some dissatisfaction with a product or service to the responsible company. A second order incentive is the opportunity to receive points that convert into monetary value, based on the length of the review and the quality of associated metadata. The mixture of incentives naturally yields variation in the data including duplicate, vacuous or offensive posts, variable renderings of named entities and inaccurate metadata necessitating post-processing that is fed by human annotation.

3.2. Annotation

Greenfield, Chan and Campbell (2016) describe their experiments in crowd-sourcing annotation to support information extraction research. They note that at least some of their workforce of Mechanical Turkers seemed to be motivated by the quality of the interface design and the desire to maintain a high approval rating as well as the monetary incentives. By focusing their system improvements on interface design they elicit higher quality data without attracting a mercenary element interested only in highly compensated work.

Poesio et al. (2016) describe *Phrase Detectives*¹, a game-with-a-purpose for collecting anaphora annotation. Players’ incentives, in addition to entertainment, are interesting source material, a variable point system, the opportunity to progress through experience levels, leaderboards, the social motivations of teaming with friends in the Facebook version and prizes awarded via a lottery and also according to performance.

The Great Language Game (GLG) asks contributors to listen to short audio clips and indicate what language is spoken. Clips are currently selected apparently at random from 80 languages so that most players are not speakers of most of the target languages. Although created in 2013, The Great Language Game (GLG) has already collected millions of judgments. The developer, Lars Yencken released a corpus of 16 million judgments collected through March 2014 though we estimate that the number collected to date is more than double that amount. GLG employs incentives of information, entertainment, competition and status. Players compete against posted high scores and can brag about their accomplishments in a forum created for contributors. The game displays Ethnologue posts for languages the player has

misidentified and players report finding the work fun. In its first year, GLG created a volume of language identification judgments significantly greater than all of the judgments created to support all of the NIST Language Recognition Evaluations since the campaign began in 1996. However, these annotations are not directly useful for LRE. Because the game relies on the ability to tell players when they have gotten an answer correct, each new judgment adds little information about a clip whose language is already known though the many judgments for each clip provide information about confusability.

3.3. Exploitation

Tyson and colleagues (2016) describe their research on automated link discovery among *About.com*² texts. Their work shows that, compared to the corporate mission of recirculating users to maximize exposure to advertising, the different motivations of content creators leads them to create fewer links than desired, a problem that the research team is now addressing through a combination of automated techniques and additional human annotation.

Eskenazi et al. (2016) describes a series of dialog system research and development efforts that have employed novel incentives such as automated access to information and the promise of an improved customer experience in real world interactions. The data resulting from the efforts naturally contain challenging levels of noise and variation in speech. Eskenazi and her colleagues at the DialRC Center have extended the notion of novel incentives to apply to the research community as well as the subjects of a study or users of a system. By offering free access to their data and dialog system and by organizing a range of outreach activities, they continue to attract researcher cycles to problems of interest to them. A recurring theme of community organized shared task challenges is that: “*optimization for lab test subjects may not reflect the outcome with real users*”.

4. Language Data Collection outside HLT

Despite the obvious benefit to HLT development, initiatives outside the HLT communities have employed novel incentives more frequently in a wider range of contexts and to greater effect. In many cases, the motivation for such collections is quite remote from HLT developers’ goals. Furthermore, neither the contributors nor the leaders of the effort may see what they do as language data collection; however we will show here that their outcomes may be extremely beneficial to research in linguistics and language technology both directly and as a model of collections that we may imitate.

4.1. Librivox

LibriVox³ creates “free public domain audiobooks” by recruiting, training and organizing volunteers who record

¹ <https://anawiki.essex.ac.uk/phrasedetectives/>

² <http://www.about.com/>

³ www.librivox.org

themselves reading literary works that are out of copyright in the US. LibriVox readers also declare their recordings to be in the public domain. As of March 25, 2016, the LibriVox catalog listed 10,185 books⁴ comprising at least 57,369 hours of read speech. Approximately 86% of all LibriVox recordings are in English. However, there is at least one hour of speech in at least 31 other languages. Figure 1 shows the growing volume of recordings by language, measured in hours of speech as indicated in the LibriVox Catalog.

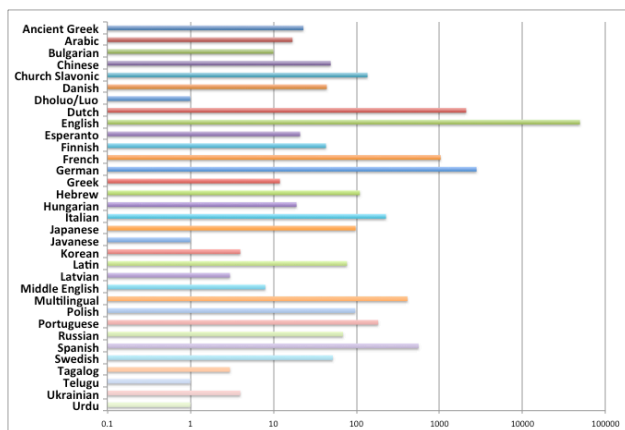


Figure 2: LibriVox Hours Recorded per Language on a log(10) scale

LibriVox recordings are typically careful readings, often of well-known works of literature for which the original written text is also available digitally. Sound quality is variable but generally good with many of the readings made in quiet environments using appropriate recording equipment and undergoing quality control by an independent producer. One or more readers may produce a single work dividing the effort either by chapter or by character. A single reader may also read multiple characters, using different voices and accents when the text seems to call for it. Most readers are amateurs from around the world, including some non-native speakers. Many LibriVox volunteers produce multiple works thus providing multiple samples of their voices over time and under different circumstances.

LibriVox recordings are relevant to a number of HLT fields including language, speaker and speech recognition. However the HLT area that makes the greatest use of LibriVox is probably speech synthesis where the large volumes of long-duration, read speech across a number of genres supplements existing data used to build TTS systems. In one early example of this use, Prahallad, Toth and Black (2007) built statistical parametric speech synthesis systems with male and female voices from a total of ~11.25 hours of LibriVox reading. They compared these to similar systems built from the Arctic Corpus (Kominek, Black 2003), designed specifically to

⁴ What the LibriVox Catalog tags as a <book> is typically a single reading of a work which could also be a pamphlet, poem or collection of poems. Additional readings of the same work receive their own Catalog record. Thus there are fewer than 10,185 unique titles.

support speech synthesis research, and concluded that “a voice could be successfully built from large multi-paragraph speech using automatic segmentation tools.” Braunschweiler, Gales and Buchholz (2010) used lightly supervised, recognition-based alignment to select paragraphs as training material for a speech synthesis system appropriate for reading longer extents of coherent text. Székely et al. (2011) experimented with approaches to clustering utterances in LibriVox readings according to voice quality parameters in order identify utterances associated with different voice characteristics and use them to build systems capable of synthesizing “speech which is rich in prosody, emotions and voice styles.” Mamiya et al. (2013) experimented with and evaluated lightly supervised VAD prior to grapheme-based alignment of LibriVox audio to corresponding text in the process of building TTS systems. The VAD system required 50 sentences of the same text to be hand aligned. To evaluate the systems they elicited 90 preference decisions from each of 20 judges who listened to system output. They concluded that the performance of the lightly supervised systems was equivalent to that of their fully supervised system. Proctor and Katsamanis (2011) elicited judgments from 13 listeners concerning the felicity of multiple LibriVox readings. Although the judges as a group clearly preferred some and dis-preferred other readers, individual preferences foiled a rigorous classification. Similarly, attempts to correlate preferences with standard prosodic measures failed to create a robust classification of reader felicity.

These studies show both the benefits of using sources like LibriVox in HLT development as well as the pre-processing needed to condition it. To the extent that the processing can be done efficiently sources like LibriVox become critical additions to the set of available LRs for HLT development, data that would be impossible to create using the traditional incentive models in our field. Each hour of recorded LibriVox audio apparently requires 2 hours of reading time and 2 to 4 hours of editing time, meaning that the initiative has elicited at least 229,476 hours of volunteer labor and probably much more. Assuming rates of \$500 per finished hour of audio, one would have paid more than \$28 million to produce the same material professionally. Volunteers make such enormous contributions for a variety of reasons. Many believe in the LibriVox mission or its connection to the broader open source or free culture movements (Erard 2007). Some enjoy reading aloud, in some cases continuing or expanding an activity they began with friends or family. Others are happy to think they are helping maintain the art of storytelling. Some clearly enjoy collaborating with others of similar interests and having the ability to control the size of their own contributions. A small number of the best readers also receive paid work through Iambik⁵, a spin-off audiobook company, or parlay their LibriVox experience into

⁵ www.iambik.com

professional narrator with Audible⁶, ACX⁷ or similar organizations. LibriVox stands not only as a data source but as a model of how initiative may use non-monetary incentives effectively.

4.2. Citizen Science: Zooniverse

Outside HLT, other research disciplines have effectively engineered environments to collect data using non-monetary incentives. Zooniverse is a citizen science portal with many opportunities to contribute to research most of which is in the hard sciences. Tasks include identifying signs of movement in star fields, classifying animal species based on photographs and transcribing museum records for insect specimen collections. The beautiful interfaces are fine grained tasking attract participants and allow them to complete meaningful tasks in minutes. More than 800,000 volunteers have registered, contributed data toward the science of many peer-reviewed publications and even made serendipitous discoveries of astronomical objects.

5. Future Directions for Language Research Development

The initiatives sketches above make it clear that there are numerous opportunities to acquire data from corpora developed under non-monetary incentives and to engineer environments with optimal combinations of incentives and workflows to develop data products for specific tasks. For example, a citizen science-of-language portal could attract equal or greater contributions because while the sciences are only one of many areas of intellectual interest, language is a common experience for nearly every human on the planet. Tasks for citizen linguists could require nothing more than native speaker ability and could scale according to the dedication of the workforce. Finally, for many, language is connected to identity so that local pride, cultural preservation and “putting ones language on the map” become additional incentives. Additionally, games-with-a-purpose, gamified interfaces and even soberer efforts that pay attention to task size and complexity relative to the workforce can increase efficiency and quality.

6. Conclusion

This paper has opened the dialog on incentives in language resource development and how they attract different workforces and require different workflows in order to optimize outcomes for a specific tasking. The HLT community is quite familiar with the impact of various monetary incentives and the effort needed to condition data acquired under non-traditional motivations, for example found data. However efforts to consciously engineer incentives and workflows within HLT have been rather limited. We described several in this paper but also believe the field needs to benchmark its data creation

efforts against external efforts that have been much more effective. Innovation in language resource creation, employing novel incentives, workforces and workflows is critical if the field is ever to seriously address the demand for HLTs for the world’s languages.

7. Acknowledgements

The author would like to thank the participants in the LREC 2016 Workshop on Novel Incentives for Collecting Data and Annotation from People. Their contributions made this paper possible.

8. References

- Braunschweiler, Norbert, M.J.F. Gales, Sabine Buchholz. 2010. Lightly supervised recognition for automatic alignment of large coherent speech recordings, Interspeech, Makuhari, Japan, September 26-30.
- Campbell, Nick. 2016. Herme & Beyond; the Collection of Natural Speech Data. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.
- Cieri, Christopher, Walt Andrews, Joseph P. Campbell, George Doddington, Jack Godfrey, Shudong Huang, Mark Liberman, Alvin Martin, Hirotaka Nakasone, Mark Przybocki, Kevin Walker. 2006. The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research, 5th International Conference on Language Resources and Evaluation, Genoa, May 22-28
- Cieri, Christopher, Linda Corson, David Graff, Kevin Walker. 2007. Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora, Interspeech: 10th International Conference on Spoken Language Processing, Antwerp, August 27-31
- DeAngelis, Tori. 2010. Found in Translation in Monitor on Psychology, 2010, Vol 41, No. 2, print version: page 52, American Psychological Association, <http://www.apa.org/monitor/2010/02/translation.aspx>
- Erard, Michael. 2007. The Wealth of LibriVox: Classic texts, amateur audiobooks, and the grand future of online peer production, Reason 39:1, p. 46.
- Eskenazi, Maxine, Sungjin Lee, Tiancheng Zhao, Ting Yao Hu, Alan W Black, Unconventional Approaches to Gathering and Sharing Resources for Spoken Dialog Research. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.
- Greenfield, Kara, Kelsey Chan, Joseph P. Campbell, A Fun and Engaging Interface for Crowdsourcing Named Entities. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.
- Harvard Humanitarian Initiative. 2011. Earthquake Relief in Haiti. <http://hhi.harvard.edu/sites/default/files/publications/earthquake-relief-in-haiti.pdf>
- International Association of Conference Interpreters. 2008. Interpreting in Zones of Crisis and War:

⁶ www.audible.com

⁷ www.acx.com

- <http://aiic.net/page/2979/interpreting-in-zones-of-crisis-and-war/lang/1>
- Kominek, John, Alan W Black. 2003. CMU ARCTIC databases for speech synthesis, CMU Technical Report CMU-LTI-03-177, Ver. 0.95, Pittsburgh, PA. Carnegie Mellon University.
- Lewis, M. Paul, Gary F. Simons, Charles D. Fennig (eds.). 2016. *Ethnologue: Languages of the World*, Nineteenth edition. Dallas, Texas: SIL International. Online version: <http://www.ethnologue.com>.
- Liberman, Mark, *Oral Histories: Linguistic Documentation as Social Media*. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.
- Mamiya, Yoshitaka, Junichi Yamagishi, Oliver Watts, Robert A.J. Clark, Simon King, and Adriana Stan. 2013. Lightly Supervised GMM VAD to Use Audiobook For Speech Synthesiser. ICASSP.
- METANET. 2010. META-NET White Paper Series: Press Release, <http://www.meta-net.eu/whitepapers/press-release-en>, accessed March 16, 2016.
- Mitsuzawa, Kensuke, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, Tomoya Mizumoto, FKC Corpus: a Japanese Corpus from New Opinion Survey Service
- Poesio, Massimo, Jon Chamberlain, Udo Kruschwitz and Chris Madge, *Novel Incentives for Phrase Detectives*. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.
- Prahalad, Kishore, Arthur R Toth, Alan W Black. 2007. Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases, Proceedings of Interspeech, Antwerp, Belgium.
- Proctor, Michael, Athanasios Katsamanis. 2011. Prosodic Characterization of Reading Styles using Audiobook Corpora, 162nd Meeting of the ASA, Thursday November 6, San Diego, CA
- Székely, Éva, João P. Cabral, Peter Cahill, Julie Carson-Berndsen. 2011. Clustering Expressive Speech Styles in Audiobooks Using Glottal Source Parameters, Interspeech.
- Tyson, Na'im, Jonathan Roberts, Jeff Allen, Matt Lipson, Evaluation of Anchor Texts for Automated Link Discovery in Semi-structured Web Documents. 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož Slovenia.

Herme & Beyond; the Collection of Natural Speech Data

Nick Campbell

Speech Communication Lab
School of Computer Science & Statistics
Trinity College Dublin, Ireland
nick@tcd.ie

Abstract

This paper describes our approach to the collection of ‘natural’ (i.e., representative) data from spoken interactions in a social setting in the context of the development (through time) of expressive speech synthesis. Over the past ten years or so, we have collected several corpora of unprompted social conversations that illustrate the ‘contact’ element of speech that was lacking in many of the corpora collected by use of a specific ‘task’ with paid participants. The paper discusses the technical and ethical issues of collecting such spoken material, and highlights some of the problems we have encountered in the processing of this much-needed data. Through the use of attractive conversational devices, we have found that natural human curiosity, and an element of social programming combine to provide us with a rich source of material that complements the task-based collections from paid informants.

Keywords: task-free dialogue, spontaneous speech, data collection techniques, ethics & incentives, the common good

1. Introduction

Developing speech synthesis technology requires the collection, annotation, and analysis of large amounts of speech data and as our knowledge of speech processes grows, alongside a phenomenal growth in computer memory, processing power, and bandwidth, we find an ever-increasing need for larger amounts of material. Gunnar Fant, perhaps the founder of computer speech synthesis, firmly understood the science of voice production and built his talking machine from first principles, with no need for a corpus of examples to replicate. Denis Klatt, in his seminal work of the 80’s studied copious spectrogram printouts of actual vocalisations to increase the naturalness of his speech output by modelling the features and dynamics that he observed in the data. Joe Olive, another pioneer of this field, used actual recordings from which he cut diphone-sized segments of speech for a more precise modelling of the information carried in the transitions between the phones. (Fant, 1970; Klatt, 1987; Olive 1980)

The 80’s saw the development of machine-learning and increased use of statistical modelling with the consequent growth of multinational initiatives for the common collection of speech material from across the world, and the development of organisations such as the LDC and COCOSDA (with ELDA/ELRA coming close on their tails) and the recommendation of specifications and tools (BLARKS) for the collection and annotation of common speech material (Krauwier 1998; Mapelli 2003). On this foundation, the present ubiquitous speech technology was built.

The situation in the present century is vastly different; and the foundation technology that grew from common experiments has become integral in mobile devices and ubiquitous human interfaces. Corporations use these technologies for their daily interaction with customers and now stream almost infinite amounts of real-world data through their systems. Deep nets have evolved to process this information on massively parallel gpu devices that make the small collections of the immediate past seem very primitive in

comparison. The world of speech processing is now split in two; those that provide actual services can find more data than they need, while those in academia or smaller industrial start-ups are left with no access to the corporate streams. In parallel, ethical issues of data collection, storage, and protection arise to a frightening degree, as the potential for abuse (or leakage to unintended recipients of confidential information) becomes a more real everyday threat. We must now find new ways to collect corpora (or learn from live streaming speech processes) that meet modern size requirements yet preserve privacy.

How did speech data collection become a threatening activity? What happened in the transition from innocent spectrogram analysis to privacy-revealing spontaneous speech collections? In the pre-Snowden age, recordings were not treated with suspicion or fear. Subjects gladly contributed; as ‘giving your voice to science’ was on a par with ‘leaving your body for medicine’ and was considered an altruistic act, not necessarily requiring payment. Some incentives were provided (Call Home collections for example (Cavanaugh 1997) offered cut-rate or free calls) but the amounts of data were relatively small and the content, even if of a personal nature, was considered privileged and not open to abuse.

2. Expressive Speech Processing

Yoshinori Sagisaka of ATR in Japan introduced the ν -talk system of non-uniform concatenative speech synthesis (Sagisaka et al 1992) based on recordings of 5000 words and 503 sentences as raw material. This was considered a large corpus at the time. The recordings were from professional announcers, people trained to produce consistent ‘standard’ pronunciations in a ‘received’ quality of voice. There were no hesitations in the readings, and no laughter or other non-speech vocalisations. The recordings were segmented by hand and strict labelling applied; the phoneme set was known, and allophonic variation was taken care of automatically as being due to phonetic context dependencies. The resulting synthesis was clear, well-

articulated, and pure-‘Tokyo’! No regional or personal deviations were allowed.

These methods of speech synthesis produced clear formal-sounding sentences (each utterance had a well-marked full-stop at the end!) suitable for announcements, broadcast-news reading, and impersonal information provision. They sounded robotic because a) the signal was manipulated, and b) the text was ‘unnatural’. But that was the nature of speech synthesis at the time. These were Talking Machines that rendered text into speech, based on the dream of reading machines from the 70s. There was no need for laughter or hesitations as these were perceived as speech ‘defects’ and the sign of an untrained speaker or an amateur performer.

With the growth in the availability of speech recordings we were able to extend ν -talk to produce CHATR and by removing the signal processing to instead use raw speech segments in unprocessed form for concatenation were able to reproduce the known voice of any given speaker (see the paper on ‘CHATR the Corpus’ in the main conference). This brought with it the sometimes embarrassing facts of ‘natural’ speech that varied from the ‘received’ dialect/accents and displayed all manner of ‘spontaneous’ speech phenomena as were found in the original recordings. Talking Machines had become capable of conversational speech.

Given that the synthesiser was now able to replicate any voice, dialect, or speaking style, the question remained as to what types (variations beyond the mean) would be required for conversational speech synthesis. Even for reading books, a considerable range of voice qualities and expressivity would be required; but for ‘interactive’ synthesis where the machine would need to replicate human characteristics of speech, the territory was uncharted. Would the machine need to laugh, for example? Would it need to cough? Singing synthesis was already being explored elsewhere as an independent field of research, and poetry-reading was perhaps too specialised a form of vocalisation to require synthesis. The limits were unknown and hence the need for representative corpora.

The conundrum here was that a well-designed corpus would produce all the examples that it was conceived to collect, but there was no specification of what that coverage might require. On the other hand, an undesigned corpus was at that time a contradiction in terms, leaving too much to chance. We now have task-specified collections where applications stream in data from countless users, but when the base technology was still under development, that was too ambitious to even consider. The Table-Talk Corpus was a first attempt at resolving this data collection problem.

2.1. Table Talk

Table-Talk (ISLRN: 545-953-122-584-3) was an early experiment in multimodal speech data collection. Five participants met over a period of three days to sit together and talk surrounded by microphones and cameras that recorded everything from several angles. No task was specified and no topic set. Here we discovered the wonderful facility that humans have for just talking (Dunbar 1998). Silence in social situations is taboo, so people sharing a common space start spontaneously to chat. No new science was involved but

the data we collected showed intriguing patterns of interaction dynamics and vocal usage. There were very few ‘well formed sentences’ among these utterances. Instead there was a rich variety of laughter and spontaneous ‘chirping’ as topics emerged and interest grew around them. Topics decayed away to be replaced by others, arising from points previously raised, or completely introducing a new subject.

This experience emboldened us to propose the Expressive Speech Collection (funded by the JST) whereby people volunteered their speech in exchange for token payments (and the possibility to keep the recording device (a mini-disk recorder) for personal use). No constraints were made on the speech to be recorded and participants were encouraged to keep the recorder active at all times so that when an interesting event occurred there would be no need to interrupt the flow by switching on the machine. Although there are strong ethical constraints on deliberately inducing fear in participants, the recording of natural fear (in the case of an earthquake for example) was considered inoffensive. In the five years we were recording there was not one fear-inducing quake, but several minor tremors. Similarly ‘joy’ and ‘surprise’ can be difficult to elicit (fake?) in the studio but are common occurrences in nature. We had faith that what was being collected would be representative of the types of vocal activity that would be needed by a speaking machine that was to operate in the real world, perhaps taking the part of a remote human in a local (and possibly translated) conversation.

2.2. JST ESP

The findings of the JST/ESP data collection (Campbell 2002) have been reported widely elsewhere. Sufficient here to note that they revealed a wealth of unexpected facts about how the voice is used in social situations in the real world. They also revealed the extent to which non-verbal information is used in place of linguistic structures, and how the social element in interaction absolutely dominates for most of the time. There were few extremes of joy or sadness but plenty of everyday expressive speech and many meaningful variations in voice-quality and speaking style. Previous ideas of how spoken interaction worked had been based on linguistic components alone and a new field of expressive interaction was opened up. Previous recordings of spoken interaction had been predominantly task-based, and the participants (being paid for their expensive time) were usually loath to digress from the specified task to ‘waste time’ in ‘mere’ social chit-chat! In this context it is interesting to note the difference between Petukhova’s PhD thesis (Petukhova and Bunt 2012) and the resulting ISO standard that arose from it ISO 24617-2 makes no reference to ‘contact events’, whereas two levels of interaction in the thesis depend on them. The ISO standard lacked evidence for social contact because the majority of the corpora that had been collected were planned top-down and specified the tasks (and therefore the coverage) of the speech in advance. No social contact occurred. The paradigm itself renders the collected speech unnatural in a social sense.

2.3. D64

From the ESP insights we gained on the value of spontaneous interaction it was a short step to the recording of the D64 Corpus in Dublin (Oertel 2010). We booked a hotel apartment for three days (number D64) and populated it with equipment and people. No money changed hands, and no instructions were given, though each participant did sign a consent form acknowledging that everything was to be recorded, warning that indiscretions were inadvisable, and giving each the right to withdraw at any time or have recordings erased from the record if so desired. Food and drink were provided (including wine on the third day!) and devices were left running from before the start to after the end (the setup and calibration of various recorders actually makes a particularly interesting part of the corpus as stresses were high given the time constraints and technical complexity of the equipment). The participants quickly became friends, sharing some extremely personal information at times, and no thought was given to forms of payment - this was fun! But the participants were all academics - and there is a general expectation in this community that effort is to be freely contributed (paper reviews for example) towards the greater good of generating knowledge

2.4. D-ANS

Perhaps this philosophy underlay the Dublin-Autonomous Nervous System Corpus of Biosignal and Multimodal Recordings of Conversational Speech (D-ANS: Hennig 2014), as the participants were members of the same lab, taking a break and chatting in front of cameras while wearing biosensors. The conversations that arose were without doubt 'natural' and completely spontaneous, and the biometric readings that we collected in addition to the audio and video data again revealed patterns of the cognitive processes underlying social speech production that were not known beforehand. This was not 'work' per se but a voluntary effort on a very small scale to increase our understanding of speech processes. The challenge now, having learnt the worth of spontaneous and informal collections is to generalise them to a larger scale and to automate the subsequent processing. A manually segmented and annotated corpus of even this small size can take several years before coming to fruition (Gilmartin et al 2013).

3. Herme & beyond

Herme was different. Here we employed one-to-one conversations instead of group talk, and we had no idea how many participants would take part (Han et al 2012).

Herme was a small motorised ©LEGO robot platform that supported a web-cam (with high quality microphone) and triggered a new conversation when a person was spotted (by use of OpenCV face recognition). The device was exhibited as part of a three-month exhibition (Human+) in the Science Gallery in Dublin; a high-tech art space where members of the public can come in from the rain to enjoy science & technology with some coffee and free wifi.

We maintained total control of the conversational flow from the start, as Herme always took the initiative, listened to any responses (without ASR) and responded with a backchannel or changed the subject according to a predetermined

sequence of conversational utterances. Both wOz and automated versions were tested but the human operator proved significantly better than the algorithm at keeping a participant interested in the conversation. The sequence of utterances was identical in both paradigms but the timing of utterance onset was too delicate a control for the software to compete. While not the focus of the current paper this aspect of timing control for dialogue speech synthesis is currently work in progress, and the data from 'failed' conversations is invaluable for training statistical models.

Natural curiosity was probably the main incentive driving most of the Herme conversations - people were attracted by the object - it moved, made noises and most importantly had a display which showed what it saw. When a person approached, their own face appeared in Herme's display, with a circle drawn round it to show that she¹ had recognised them as a person. When Herme spoke at that point it was not immediately natural for people (as observers) to respond, but when she repeated the greeting most people responded with a greeting in return - accepting on the second utterance that the robot was talking to them and becoming active participants.

The voice of the robot was childlike, and the childlike innocence (and directness) of the questions she asked had an appeal that many people instinctively responded to - and answered politely (or jokingly) in response. Most participants stayed for about three minutes, the length of a complete conversation, and then signed a release form giving researchers permission to use the data when asked to do so by the robot. It was clearly signposted that all conversations were being recorded. Over the three-month period more than 1500 people voluntarily took part in a conversation with the robot and about two-thirds signed the consent forms to allow us use of their data.

Gilmartin & Su (forthcoming) have recently extended Herme to produce 'Cara', a conversational autonomous relational agent, which was recently exhibited² as part of the All Ireland Linguistic Olympiad at Trinity College in Dublin. This software instantiates a full dialogue system and uses ASR in conjunction with Voice-Activity Detection to inform the dialogue manager of which utterance to render next and at what time. The Olympiad attracts some of the brightest and most inquisitive of Irish schoolchildren to compete on linguistic puzzles and our side-exhibition provided a rich source of interaction behaviour as the children took turns to chat with the robot during their breaks.

The experience was mutually beneficial - the curiosity of the children prompted them to test the limits of the robot's dialogue capabilities, providing a learning experience for both sides, and fun for the participants while producing invaluable data for the developers. Of course the system failed often - the state of the art in autonomous dialogue systems is still far from ideal, but from the point of view of research, if everything runs smoothly then there is little left to learn, and as our goal in collecting these data is to gain experience, then failure (of a dialogue) is as valuable to us as 'success'.

¹Herme is generally thought of as 'female'

²mid-March 2016

4. Discussion; Generalising the Process

A complaint from an industry representative at a recent Interspeech lunch was that many of the scientific papers were reporting results from corpora of less than 20-hours of speech material, pointing out that results from such small studies just don't generalise to be useful for solving real-world problems. He might have said 200-hours, the point would have been the same.

Corporate analysis of speech data reported at a recent ICASSP cited 200,000 hours of speech material as normal for training. The major service providers have solved the data collection problem and are now tackling the issues of working with really 'big'-data but are unable for a variety of reasons to make that resource available to a wider public. Nor do they perhaps see the need to solve some of the problems that academic researchers find interesting.

Fortunately many corporations take in interns for short periods and experiments can be made (under strict limitations of confidentiality) on in-house data ("this call may be recorded for training purposes") the general results of which can be published more widely.

Social media also provide rich streams of interesting material but apart from the technical and legal problems with tapping these sources, the 'language' they use is perhaps unique to the medium. It may be evolving to form a common subset of human language with its own grammar and syntax (hash-tags, etc.,) but is less useful for synthesis.

The need for task-based conversational data can presumably be satisfied by the applications that provide the services that meet the tasks, but there is still a need for non-task-based, primarily social speech data for the next generation of human-machine interfaces. Machines may not need to replicate the full range of human sounds in synthesised speech but they will, we argue, be required to process this information to make inferences about the human cognitive states in an interaction so that an appropriate response may be served by the machine.

5. Conclusion

This paper has described our approach to the collection and analysis of speech data for the development of interactive speech synthesis for use in dialogue systems. We firmly believe that it is of more value to collect unstructured data that yields fresh knowledge on speech processes and that the top-down design constraints of a 'well-designed' corpus can prevent these spontaneous natural features from emerging. As our systems develop, so we can use them to collect more material. The element of fun in interacting with a machine in a very human way seems to motivate people to help us, and we learn much from what they try to make the machine do. The types of voice, speaking-style, and vocal activity have surprised us in the ways they deviate from standard descriptions of linguistic use. We infer that the linguistic models, and the types of speech that synthesisers are generally trained on are abstracted away from the complex details of everyday performance and encapsulate instead a higher knowledge about the language and speech per se, rather than an encoding of actual everyday performance. The value of collecting data in the wild far exceeds any financial or other costs and will, we hope, help us to provide

an interface that is more in touch with the actual everyday needs and expectations of the people who will have to use this technology in speaking devices of the future.

Acknowledgements

This work has been carried out at ATR in Japan and in the Speech Communication Lab in Dublin with funding support from JST, Kaken, SFI, CNGL, and the ADAPT Centre. The work owes much to the many inspirational researchers who have contributed time and effort and helped in this great learning experience.

References

- Fant, G., (1970) *Acoustic Theory of Speech Production*. Mouton De Gruyter. ISBN 90-279-1600-4
- Klatt, D., (1987) "Review of text-to-speech conversion for English" *J. Acous. Soc. Amer.* 82, 737-793
- Olive, J., (1998) "A scheme for concatenating units for speech synthesis", in *Proc Acoustics, Speech, and Signal Processing, IEEE International Conference ICASSP '80*. (Volume:5) Apr 1980, pp.568 - 571
- Krauwier, S., (1998) "ELNET and ELRA: A common past and a common future", in *ELRA Newsletter Vol. 3 N. 2*.
- Mapelli, V., Choukri, K., "Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps", *ENABLER project internal report, Deliverable 5.1, 2003*.
- Sagisaka, Y., Kaiki, N., Iwahashi, N., Mimura, K. (1992) "ATR nu-talk speech synthesis system". *Proceedings, International Conference on Spoken Language Processing*.
- Canavan, A., David G., and Zipperlen, G. (1997) "CALL-HOME American English Speech" LDC97S42. DVD. Philadelphia: Linguistic Data Consortium.
- Petukhova, V. and H. Bunt (2012) The coding and annotation of multimodal dialogue acts. In *Proceedings 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul.
- Oertel, C., Cummins, F., Edlund, J., Wagner, P., and Campbell, N. (2010) "D64: A corpus of richly recorded conversational interaction". *Journal on Multimodal User Interfaces*, pages 1 – 10.
- Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*. Harvard Univ Press.
- Campbell, N., (2002) "The Recording of Emotional speech; JST/CREST database research", in *Proc Language Resources and Evaluation Conference (LREC)*.
- Gilmartin, E., Hennig, S., Chellali, R., and Campbell, N. (2013). *Exploring sounded and silent laughter in multi-party social interaction - audio, video and biometric signals*. Valetta, Malta, October.
- Hennig, S., Chellali, R., and Campbell, N. (2014). *The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech*. Reykjavik, Iceland.
- Han, J.G. et al. (2012) "Speech & Multimodal Resources: the Herme Database of Spontaneous Multimodal Human-Robot Dialogues". *8th LREC, Istanbul, Turkey, 23-25 May*.

FKC Corpus: a Japanese Corpus from New Opinion Survey Service

Kensuke Mitsuzawa[†], Maito Tauchi[†],
Mathieu Domoulin[†], Masanori Nakashima[†], Tomoya Mizumoto[‡]

[†]Fuman Kaitori Center

[†]6-5-1 Nishi-Shinjuku, Shinjuku-ku, Tokyo 163-1333, Japan

[‡]Tohoku University

[‡]6-6-05 Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi 980-8579, Japan

[†] {kensuke_mitsuzawa, maito_tauchi, domoulin_mathieu, masanori_nakashima}@fumankaitori.com

[‡] tomoya-m@ecei.tohoku.ac.jp

Abstract

In this paper, we present the FKC corpus which is from Fuman Kaitori Center (FKC). The FKC is a Japanese consumer opinion data collection and analysis service. The main advantage of the FKC is the system that awards greater points to user input containing more information, which encourages users to input categorical information. Thanks to this system, the FKC corpus has consumers' opinions with abundant category and user demographics, and is considered to serve multiple NLP tasks: opinion mining, document classification, author inferring and sentiment classification. The FKC corpus consists of 254,683 posts coming from 25,092 users. All posts are checked by annotators who are working for the FKC in crowdsourcing. The posts in the FKC corpus mainly comes from mobile devices, and one third of them are about products or events related to daily life. We also show some correlations between point incentive and users' motivation which keeps posting their opinions with abundant category information.

The FKC corpus is available under an original license of the FKC. Currently, the FKC gives permission to use directly, thus, those who hopes to use the FKC corpus needs to send request to first author.

Keywords: Social Media, Corpus construction, Crowdsourcing

1. Introduction

Public datasets extracted from the web are a popular data resource for NLP research. This is especially true for modern NLP research which makes increasing use of machine learning for such research applications as document classification (Boley et al., 1999; Schenker, 2003), sentiment classification (Zhang et al., 2015), opinion mining (Ori-maye et al., 2012), and author inferring (Mukherjee and Liu, 2010; Nguyen et al., 2011).

There are several issues that researchers commonly face when using many of the public datasets made of information on the web. First, these datasets are often noisy. They require time-consuming pre-processing before they can be used. Second, these data resources tend to be lack in contextual information (i.e. metadata) such as author profile, likewise class metadata can be inconsistent. Thus, analysts or researchers must often manually label their data before use, as in (Noll and Meinel, 2008).

In this paper, we introduce a novel Japanese language corpus. This is extracted from data accumulated by *Fuman Kaitori Center* (FKC)¹, which is a Japanese consumer opinion data collection and analysis service opened in 2015. “*Fuman*” means dissatisfaction in Japanese. The core concept of the FKC is to collect consumers' negative opinions about companies and their products or their services in exchange for a small monetary reward. This monetary reward is exchangeable with a gift card which is able to be used in an electronic commerce service. As a running web service, the FKC is accumulating data at the rate of 5-10,000 posts a day as of mid 2015. On the business side, the FKC offers an analytics dashboard and custom reports to whom wishes to know opinions on specific products or services as shown in Figure 1.



Figure 1: Analytics service from the FKC. Business users are able to check latest statistics with an Analytics dashboard (Left), to check suggestions from data with an Analysis Report (Right)

Considering the FKC corpus as dataset for NLP tasks, the FKC corpus has several major advantages. First, the corpus includes metadata such as user profile information accompanying the posts' textual content. In addition, the corpus is less noisy than other comparable public datasets, and the corpus is more focused, only including relatively short, negative opinions. Secondly, the FKC corpus is collected from a live service and is thus growing every month, making possible research applications that require time-series data. We believe that the FKC corpus can be a useful data source for a great variety of NLP tasks.

In this paper, we first show related dataset and platforms in Section 2. Next, a brief introduction of the FKC is presented in Section 3. The Section 4 describes statistics in the

¹<http://www.fumankaitori.com>

FKC corpus. The Section 5 shows correlations between a point incentive system of the FKC and users' motivation. We give some examples of NLP application in Section 6. Finally, we make the conclusion in Section 7.

2. Similar datasets for NLP tasks

2.1. Twitter

Twitter is a global-scale SNS service used by people for sharing short thoughts, opinions, and observations in near-real-time either publicly or to a private group of "followers". For several years now, Twitter has been a popular resource for NLP-related research (Sasa et al., 2010; Pak and Paroubek, 2010). But using text data extracted from Twitter causes some problems. For example, it is hard to classify tweets by their topics, moreover user demographic information tends to be unknown. To be fair, there is some metadata in Twitter, like user's age and location for user demographic information, and hashtags and geo-tagging for tweets. But, as we mentioned, the user demographic information tend to be unknown because there are less merits to fill in for ordinary users. Filtering on hashtags might miss relevant posts without hashtag, otherwise that might include unrelated posts that have the hashtag spuriously, making the dataset potentially very noisy. Therefore, it is laborious task to make clean data from Twitter.

While the FKC corpus is significantly smaller than data extracted from Twitter, it is more focused, with more well-defined categories and topics. Moreover, the FKC corpus has user profile information which adds demographics to the analysis.

2.2. Youtube

Youtube allows its users to post comments for each video. These comments can be used as a relevant data source for such tasks as opinion mining. For example, Uryupina et al. (2014) uses the posted comments from promotional videos as a dataset for opinion mining. While their dataset includes some metadata, such as the video URLs and external links to related products, it does not include any user profile information. In addition, the comments are not categorized, therefore it includes some irrelevant comments, making the dataset rather noisy.

Compared this dataset with the FKC corpus, it has the advantages of user profile and less noisiness.

2.3. Rakuten Data

Rakuten, which is one of the largest e-commerce company in Japan, makes several dataset available². The one of their dataset, the Rakuten Ichiba dataset includes product data for over 150 million items as well as over 64 million user reviews about these items. Moreover, it is notable in a lot of metadata, such as user profile and review rating.

While the Rakuten dataset has review text and a lot of metadata, their reviews are limited to a specific domain. For example, reviews in Rakuten Ichiba data are only for products, also for shop owners who sell products inside Rakuten Ichiba. On the other hand, the FKC collects opinions without domain limitation such as "human relationship", "pub-

lic service" and "politics", which are useful for analysts or researchers who carry out public opinion analysis.

2.4. MPQA opinion corpus

MPQA opinion corpus is annotated dataset which is consisted of 506 documents mainly from news articles. This dataset is open at website³ and dataset description is in Wiebe and Theresa Wilson (2005).

MPQA dataset is worthy because of its wide variety of metadata information. In this dataset, *private state* (Ex. emotion, sentiment, belief, speculations etc.) metadata is annotated for words and phrases. Moreover, metadata is categorized by its expression level which is from direct expression to indirect expression. And the text is well-formatted style because its documents are mainly from news articles.

Although MPQA is good for its rich annotations, the dataset contains static information, from which public opinion is difficult to determine. The FKC corpus comes from lively posts, thus we are able to know public opinions from it.

3. A brief introduction of FKC

Fuman is a Japanese word which is usually translated into English as discontent or dissatisfaction. It can be tied to various negative feelings such as anger, sadness, disappointment, frustration and so on. Most kinds of *fuman* are posted to the FKC by consumers when they are faced with a recent unsatisfactory experience from a product, service or company.

We provide consumers' opinions to those who seek them for purposes of improving quality of service or products. Thus, the FKC is a way for consumers to communicate indirectly with the company they are dissatisfied about, and hopefully lead to an improvement in the situation. This is indeed the business model of the FKC, which makes money by selling access to valuable consumer opinion data to interested companies. To realize this concept, the FKC has been collecting user opinions since March 2015.

Consumers must register on the FKC service via its mobile application (iOS and Android) or its website. The registration form is a simple and can be filled by anyone who is capable of reading Japanese at a basic level. Figure 2 shows main functions in the FKC service. Users of the FKC can post their negative opinions from simple page (Right), also they can watch posts coming from other FKC users (Left). The FKC rewards users with points in return for their posts. Once registered, users can post their opinions. Table 1 shows the schema of the post in our corpus. All the metadata fields are optional in order to simplify the post process as much as possible.

Point value grows with the opinion's quality (the length of the post, and other criteria). The point also increases as a user adds metadata relevant to the post (adding category, product/service name, company name). Table 2 shows the schema for the user profile. Most of the user profile in-

²<http://rit.rakuten.co.jp/opendata.html>

³<http://mpqa.cs.pitt.edu/>

Table 1: Contents to be posted as fuman

Field	Essentiality	Data type	Example (English translation)
fuman	mandatory	free text	電車が毎日、遅延してばかり (Train is behind schedule everyday)
proposed idea for fuman	optional	free text	余裕をもったダイヤにした方がいい。(Train company should adjust a timetable)
target of fuman	optional	free text	東京線 (TokyoLine)
service provider of target	optional	free text	東京鉄道 (TokyoRailway)
sub-industry	optional	categorical	駅・電車 (Station & Train)
industry	optional	categorical	公共・環境 (Public Service)

Table 2: User profile

Field	Essentiality	Data type	Example (English translation)
gender	optional	categorical	男性 (male)
birth year	optional	integer	1990
job	optional	categorical	会社員 (employee)
state	optional	categorical	東京 (Tokyo)



Figure 2: Main function of the FKC service. FKC users can post their opinions with a page for posting (Right), they can watch posts from other FKC users with time-line (Left)

formation is also optional to ease the registration process⁴. Thus, a post containing only a short sentence and with no additional optional fields set has the lowest value. The maximum price can only be reached by a quality post with all optional fields filled-in for a user who has filled in all their own personal information. This system promotes user to fill user profile.

3.1. Point and procedures for exchanging

As we mentioned above, the point in return of their posts has real monetary value, which is exchangeable with gift cards for an electronic commerce service. Mostly, this point is from 1 to 10 for a post, about 5 on average. As of March 2016, 1 point is always equal with 1 Japanese Yen. In Japan, a bottle of mineral water or a can of coke is around 100 Yen. Thus, around 20 posts have almost same value of

⁴Putting user profile is mandatory from December 2015 to collect more precise opinions and to know sender of opinions more precisely.

a soft drink.

As of March 2016, the FKC is providing only *Amazon.co.jp gift card*[®] as an exchangeable gift card. FKC users are able to ask the FKC to exchange their points with the gift cards. There are 2 advantages to use the Amazon gift card. First, Amazon.co.jp is one of the most popular electronic commerce service in Japan, therefore, FKC users are able to purchase everything with the gift cards they get. Second, FKC users can receive a code of gift cards by e-mail, which makes procedures easy. The FKC sets 500 Japanese Yen as the minimum value of exchange, therefore, FKC users must accumulate at least 500 points to ask to exchange.

All exchange procedures are done via Internet. First, FKC user sends a request to exchange through applications of the FKC service. Second, the FKC reduces the point from accumulated user's point, and send a gift code of Amazon.co.jp gift card with e-mail. Finally, the user is able to use it.

3.2. Metadata in post and user profile

3.2.1. Metadata of post

The first metadata field is the *industry* and *sub-industry* of the company that the user post is about. Sub-industry is a sub category within the broader industry category. Table 1 shows an example whose industry is "Public Service" and sub-industry is "Station & Train". We have 14 industry categories and 10-13 sub-industry for each.

There is also a company/organization field, and a product/service name field that the user can either select from the existing list, or enter if there is not in our database yet.

3.2.2. Metadata of user profile

FKC users can register following 4 user profile information. The *None* are recorded in fields if a user does not choose.

Gender The gender is a categorical value which can be either male, female.

Prefecture (State) The area of residence, as a categorical value which can be set to any one of the 47 prefectures of Japan.

Birth year The birth year is a 4 digit integer.

Occupation (Job) The main occupation of the user. We set 12 typical occupations in Japan.

- 経営者・役員 [owner/board member]
- 会社員（事務系） [employee (office worker)]
- 会社員（技術系） [employee (engineer)]
- 会社員（その他） [employee (else)]
- 専業主婦（主夫） [housemaker]
- 専業主婦（主夫） [housemaker]
- 学生 [student]
- 公務員 [public employee]
- 無職 [no job]
- パート・アルバイト [part time]
- 自営業 [self-employed]
- その他 [else]

3.3. Annotation

An important feature of the FKC is that anyone can register and post anything as the content of their posts. Therefore, there will inevitably be some undesirable posts. To cope with such posts, native Japanese speaking operators manually annotate posts. They carry out three kinds of annotations; 1: label posts with a content-check flag, 2: correct category mistakes, 3: normalize the company and product name fields. For 2 and 3, we save *None* if a users does not choose them.

3.3.1. Filtering out unsuitable posts

All posts are assigned “content-check flag” label, which identifies a post as good or bad. A good post means it can get points, whereas a bad post should not result in any point reward. Since points given for posts have real cash value, there is a real business benefit for filtering out posts that are gibberish, incomplete/meaningless, uninformative or that use offensive words. Mainly, we give bad content-check flags by following reasons,

Duplication The post is completely same or extremely similar to already posted one.

No meaning sentence The post which has no meaning as Japanese is given bad flag. For example, “ああああああ (aaaaaa)”.

Positive opinion We give bad flag when a post means positive opinion, and has no negative opinion at all. For example, “きのう食べたカレーはとても美味しかった。あしたも食べたい！（It was delicious curry I ate yesterday. I would like to eat tomorrow!”)

Offensive Posts containing personal information or those that are offensive are marked as bad including those mentioning untitled civilians or containing racial discrimination or abusive words⁵.

For example, “スーパー店員の山田太郎という店員の感じが悪かった。（A shopkeeper named Taro Yamada, he was disgusting.)”⁶.

3.3.2. Correcting mis-categorized posts

It can unfortunately happen that users do not select the correct industry/sub-industry category given the content of their post. The operators check these categories and correct mistakes when they are found.

For example, “Public Service” is correct category in an example of Table 1. But some users might post their opinions as “Sightseeing & Leisure” if their posts’ context is like “Negative opinions for passengers in a train when I was on the way to leisure places”. In such case, the operators correct “Sightseeing & Leisure” to “Public Service”.

3.3.3. Normalization of free-text fields

The company and product name fields allow direct user input. Since users can enter the same entity in multiple forms, this field must be normalized. For example, a user might mention “Apple Inc.” as “apple computer” or “vendor of iphone”, which neither is the actual name of the company. To cope with such ambiguity problem, our operators manually normalize the data to an agreed upon single value, “Apple Inc.” in this case.

For now, we are carrying out normalization only for the “company/organization” field because “product/service name” field has such a sheer variety of products and services that users may be referring to that it is hard for operators to cope with all of them. The procedure for normalization is following:

1. We made a master database of representative manufacturing and hospitality companies in Japan. This is because most users mentioned about company.
2. We make relationship between master data and values in “company/organization” that user mentioned. If the master data does not have “company/organization”, we clean up the text and add it into the master.

3.3.4. Annotation procedure

For annotation, we hired 8-10 part-time workers as annotation operators. Each post is annotated by only one worker. We put a priority on speed. The FKC is running platform and new posts are continuously being created⁷. Point reward must be done with a reasonably short delay for best customer service.

Given that each post is reviewed by only one part-time worker, we asked one of our employee, also a native

⁵We removed posts which are categorized into Offensive from the FKC corpus because this category includes sensitive contents.

⁶This sentence is just a fictional example. *Taro Yamada* is a common fictional name in Japan as same as *John Smith* or *John Doe* in American culture.

⁷As of September 2015, the FKC gets an average of close to 10,000 posts a day.

Japanese speaker, to double-check the annotations of the part-time workers. As this employee knows our rules well, we believe this system is enough to ensure the overall quality and accuracy of the annotations.

To reduce mis-annotations as much as possible, we run training sessions and our employee provides feedback. During the training phase, our employee explains annotation rules to part-time workers who then apply the annotation procedures to 1,000 posts. When they finish their annotation tasks, our employee checks mis-annotated posts and lets them know their mistakes in detail as feedback. Finally, we ask them to annotate incoming posts. Even if after training, our employee gives feedbacks to them if they make mistakes.

We recruited them in some ways; from SNS like Twitter or Facebook, introductions from our employees’ families or friends. Some part-time workers live near from our office, others far from our office. Considering this situation, we asked them to work at their home for the purpose of making our procedure in uniform. Thus, all training and feedbacks are carried out with *Skype*[®] which is online conversation tool.

As a result of this training and feedback efforts, we have high agreement rate on “content-check flag” and correction of mis-categorized posts between our employee and part-time workers. We have 99.5% averaged agreement rate on all part-time workers for “content-check flag”. And we have 99.2% on mis-categorized “industry” category and 99.0% on mis-categorized “sub-industry”.

3.4. Data format

Our corpus is provided with JSON format data as shown in the upper part of Figure 3. The JSON format is easily converted into XML format because we put a script with the corpus. In this data, every item has post-meta-data and user-meta-data. The file size is around 180 MB with JSON format.

3.5. Corpus License

The FKC corpus is now available under an original license of the FKC, and is only for research purpose. Currently, our license is available only in Japan. We are working to make the FKC corpus available also for researchers in overseas. To use the FKC corpus, the one have to make a contract with the FKC directly. The First author is helping to give permission to researchers. Those who hopes to use the FKC corpus needs to send e-mail to first author and ask permission to use.

4. Corpus statistics per device and user demographics

In our corpus, there are 254,683 posts and 25,092 users. Table 3⁸ shows some basic statistics about the devices were used to write posts on the FKC by its users⁹. The “others” category includes minor mobile devices as well as unknown devices. Most posts are from Android and iPhone mobile devices, with an almost 80% share of posts. The average

⁸Statistics collected from tokenized posts using MeCab 0.996.

⁹Information parsed from User-Agent string

```

{
  "normalized_company_name": "市役所",
  "product_category": "地方行政",
  "user_number": 1,
  "fuman": "収入がゼロでも徴収される糞制度。",
  "state": null,
  "product_name": "国民健康保険",
  "birth_year": null,
  "status": "ANNOTATED",
  "company": "市役所",
  "job": null,
  "gender": null,
  "industry": "政治・行政",
  "proposals": null,
  "time": "2015-03-18 22:35:42"
},
{
  "normalized_company_name": null,
}
<?xml version="1.0" encoding="UTF-8" ?>
<0>
<normalized_company_name>市役所</normalized_company_name>
<product_category>地方行政</product_category>
<user_number>1</user_number>
<fuman>収入がゼロでも徴収される糞制度。</fuman>
<state /><product_name>国民健康保険</product_name>
<birth_year /><status>ANNOTATED</status>
<company>市役所</company>
<job /><gender /><industry>政治・行政</industry>
<proposals /><time>2015-03-18 22:35:42</time>
</0>
<1>
<normalized_company_name /><product_category>その他</

```

Figure 3: Data format example of JSON (Up) and XML (Down)

Table 3: Statistics per device

device	#post	Avg.tokens	Avg.character
Android	102,378	26.867	46.734
iPhone	97,081	27.298	47.563
PC	48,372	34.404	59.346
iPad	5,436	20.788	50.075
others	1,416	27.207	50.995
total	254,683	28.608	49.692

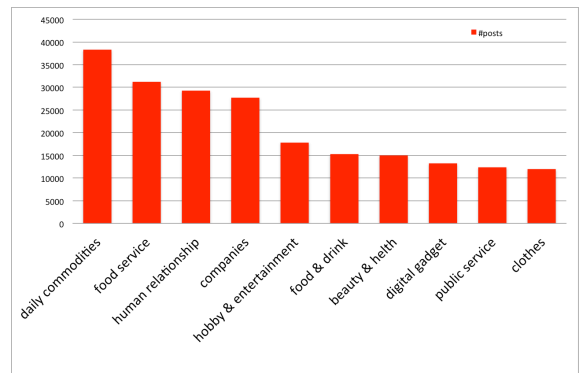


Figure 4: Top10 industry ranking. x axis is category name in industry attribute, y axis is #post for it.

character length of posts made on these two mobile platforms is 46-47 characters. Compared with posts from PC,

Table 4: Statistics of content-check flag

content-check flag	#post	ratio
Good	241,678	0.948
Bad	13,005	0.052

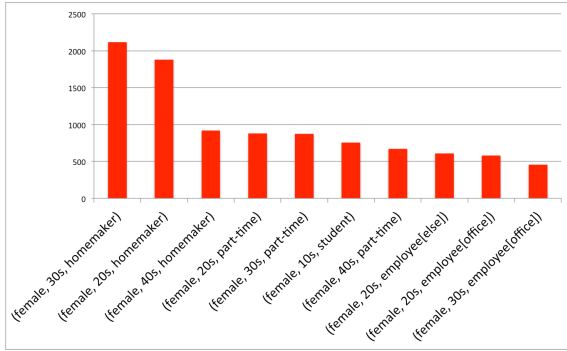


Figure 5: Top10 user demographic of (gender, age, job). x axis is (gender, age, job) and y axis is #user for it.

posts from mobile devices are shorter about 10 averaged character length as well as averaged token length. It is assumed that users using mobile devices tend to express their opinions more briefly than PC users. According to Neubig and Duh (2013), the average character length among Japanese Tweets is 40-45. From this observations, we can say that posts in the FKC mostly came from mobile devices, and the its post length is close to Twitter.

Annotation operators have labeled the posts as good or bad and this information is stored in the “content-check flag” as in Section 3.3.1. Good posts far outnumber bad ones, by a ratio of almost 19 to 1, as 241,678 posts (about 95% of all) are good posts, and only 13,005 (about 5% of all) posts are bad. From this observation, we can say that the vast majority of FKC users follow the guidelines about writing good posts.

As for the “industry” and “sub-industry” fields in Section 3.2.1., 99% of posts have a “industry” category, and 96% have both of “industry” and “sub-industry”. Figure 4 shows top 10 for posted industry categories. The top 3 categories count for as much as 39% of all posts. Considering the target of FKC users are ordinary Japanese consumers, such a selection of categories make sense, as they are such a common part of everyday life experience.

Figure 5 shows the top 10 for user demographics for the combination of gender, age and job. These top 10 combinations occupy about 38% of all users. This is a zipfian-like distribution where a few combinations are very common, followed by a long tail of all the remaining possible combinations. The post “industry” distribution for this top 10 group of user segments is almost entirely about “daily commodities”, “human relationships” and “food service industry”, mirroring the distribution of the whole dataset, meaning they are a good representative sample of all users of the FKC.

5. Correlation between user-meta-data and users’ motivation to FKC

To make clear how a system of the FKC point incentive works on users’ motivation for posting their opinions, this is shown by the relationship between tendency of filling in user-meta-data and users’ posts. The FKC service lets FKC users know the point incentive system of the FKC when

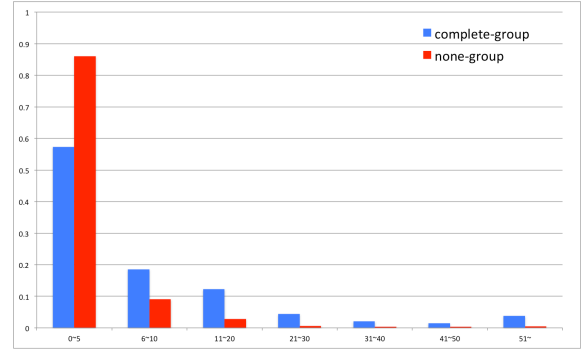


Figure 6: Distribution of #post for complete-group and none-group. x axis is segment of #post, y axis is ratio per group

new users start to use FKC services. FKC users know that points grow up when they post their opinions, so that it is considered that motivated users fill in all user-meta-data and they post much with abundant post-meta-data.

In the FKC corpus, there are 3 types of users in the point of profile information. We call users who filled all user-meta-data as *complete-group*, users who do not put any of user-meta-data as *incomplete-group*, and users who have no user-meta-data at all as *none-group*. In the corpus, 66% of users is complete-group, 20% is incomplete-group, 14% is none-group.

We investigate the correlation with the number of post, filled ratio of post-meta-data, persistency ratio of posts. For this investigation, we omit incomplete-group because it is considered that users in incomplete-group understand the FKC incentive system, however, they still refuse to put all their user profile information by any reasons. Thus, we compare complete-group with none-group in 3 investigations. In all of 3 investigations, we observe positive tendency for the FKC service.

5.1. Correlation with the number of post

If users in complete-group are motivated by the FKC point incentive system, they post much than users in none-group do. On average, users in complete-group prove to be much more prolific than users in none-group. In fact, complete-group users post an average of 11.99 posts compared with none-group users who only post an average of 3.51 posts each. Figure 6 presents distribution of #post for complete-group and none-group. We observe that the post ratio of complete-group is high in segments of much posts (all segments in more than 6 posts) compared with none-group. The ratio of users who post more than 50 posts is 3% in complete-group, by contrast, 0.4% in none-group.

5.2. Correlation with persistency ratio of posts

We show correlation between user-meta-data and the number of post in Section 5.1, however, there is possibility that some users are just new to the FKC and their posts are still a few. Considering this possibility, we might not say correct correlation from the ratio. Thus, we investigate users’ continuity of posts. If the FKC point incentive works as motivation to users, it is presumed that they keep posting their opinions to the FKC to accumulate the FKC point. It

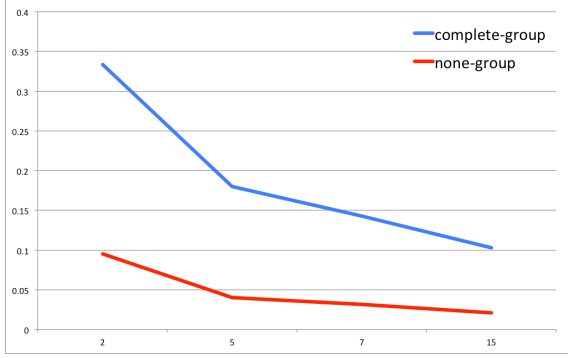


Figure 7: Persistency ratio of complete-group and none-group. x axis is k (days after first post) and y axis is persistency ratio

is desirable for the FKC that users keep posting their opinions because we are able to construct users’ models based on users’ demographic information if the FKC has enough posts coming from each user.

We define *persistency ratio* to present how much FKC user keep posting their opinions after their first posts. Here, for a user $u \in U$, we call the first date when u posted first opinion as $t_{u,0}$. With $t_{u,k}$, we count the ratio which u posted in the $t_{u,k}$ day. Still, there is possibility that u did not post his opinion in the just $t_{u,k}$ day. So, we use an adjustment parameter α to denote before and behind $t_{u,k}$ day. With the α parameter, we can check whether u posted his opinion in the range $range_{t_{u,k}}$: $[t_{u,k} - \alpha, t_{u,k} + \alpha]$ or not. The persistency ratio is defined with the following formula.

$$Persistency\ ratio = \frac{\sum_{u \in U} count\ post(u, k, \alpha)}{|U|}$$

$$count\ post(u, k, \alpha) = \begin{cases} 1 & \text{if } u \text{ post in } range_{t_{u,k}} \\ 0 & \text{else} \end{cases}$$

where

- $range_{t_{u,k}}$: $[t_{u,k} - \alpha, t_{u,k} + \alpha]$
- $|U|$: the number of users

Figure 7 shows the persistency ratio when k of $t_{u,k}$ is 2, 5, 8, 15. We use $\alpha = 1$ when k is from 2 to 8, $\alpha = 2$ when k is 15. As shown in the Figure 7, even though the difference in persistency ratio between complete-group and none-group is shrinking as k increases, there is always 2-3 times difference. From this tendency, there is clear correlation between persistency ratio and user-meta-data. We can conjecture that users in complete-group tend to be well motivated with the FKC point incentive system, therefore, they keep posting than none-group which is less motivated group.

5.3. Correlation with filled-in ratio of post-meta-data

If users in complete-group are motivated by the FKC point incentive system, we can assume that they put more post-meta-data to get more points. In other words, users in complete-group tend to have less *None* value in their posts.

#None in post-meta-data	complete-group	none-group
0	19%	17.8%
1	25.9%	23.2%
2	29.9%	26.1%
3	24.7%	29.4%
4	0.5%	2.5%

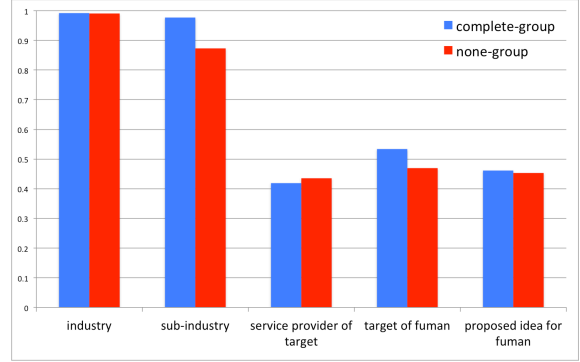


Figure 8: Input ratio of post-meta-data. x axis is attribute of post-meta-data, y axis is input ratio of it

Table 5 shows distributions of #None in post-meta-data. The complete-group has much ratio when #None in post-meta-data is from 0 to 2 compared with none-group. Interestingly, the most highest ratio in complete-group is when #None in post-meta-data is 2, by contract, the one in none-group is when #None in post-meta-data is 3. This means there is tendency that users in complete-group put plus one metadata than users in none-group.

Figure 8 shows input ratio of post-meta-data. For both of complete-group and none-group, there is a tendency that “service provider of target”, “target of fuman” and “proposed idea for fuman” are small ratio than others. These meta information are detailed information. Therefore, it is presumed that users do not remember such detailed information and skipped filling in.

The difference between complete-group and none-group is mainly in “sub-industry” (10% difference) and “target of fuman” (6.5% difference). We infer that users’ proficiency level relates to this difference. Considering 10 - 15 “sub-industry” categories per one “industry” category, users need to comprehend category structures to fill in. Also for “target of fuman”, users are required to remember product or service names to fill in. Even though users need to understand well to fill in these post-meta-data, we suppose that the FKC incentive system works as motivation for filling in.

6. Applications for NLP tasks

Many NLP tasks can make good use of the FKC corpus. The abundance of user profiles in the FKC corpus makes it especially suited to the author inferring task. One example is Nguyen et al. (2011), where they use blog corpus to construct models with the objective of predicting the author’s age. Mukherjee and Liu (2010) targets gender prediction, also from a corpus sourced from blogs. The FKC corpus can be a useful corpus to support both of these targets, as its

user profiles include both age and gender information. The FKC corpus has also other features, such as users' state, job and posts' industry categories, which are high-potential effective features.

Domain Adaptation is a task which trains a model on labeled-corpus to predict labels for other unlabeled-corpus. Dai et al. (2007) proposed domain adaptation metric between similar dataset, and Xiao et al. (2013) proposed a model between not-very similar documents such as news text and product reviews. The FKC corpus is useful again as a labeled training data for such domain adaptation models because the corpus has industry and sub-industry category for almost all posts, and there are various industry categories as in Figure 4.

7. Conclusion and Future work

In this paper, we have presented a new corpus that is consisted with lively coming negative opinions. This corpus is useful for various kinds of NLP research and we have presented some NLP metrics in which our corpus is applicable. This corpus is useful in following point: First, all posts are from ordinary consumers, which is valid data-source of opinion mining. Second, this corpus has rich metadata, which is essential information for supervised machine learning methods. Third, this corpus is less noisy compared with existing datasets of SNS because the corpus contains only negative opinions.

We showed some correlations between an incentive system of the FKC and users' motivation to keep posting their opinions with much metadata. Even though we observe positive tendency between user-meta-data and users' motivation, however, it is hard to assert causal relation clearly. We are not able to investigate how the FKC point incentive system (point incentive from user profile) works on users' behaviors because the FKC does not save all log that users changed their user-meta-data in FKC service. Besides, it is hard to conduct this analysis with the current FKC service because putting user-meta-data is mandatory from December 2015 to collect more precise opinion and to know sender of opinions more precisely. Therefore, we are planning to investigate users' behaviors via questionnaire survey, like "how do they feel about the FKC incentive system?" or "have you ever tried any of questionnaire or survey service with incentive?"

In the near future, we will publish a new version with more posts. And we will extend data input method and metadata on it. We are currently working on a new system which accepts post without registration. With this system, new posts from wide variety of users will be increased. And we believe that new metadata will lead to new applications of machine learning methods.

8. References

- Boley, D., Gini, M., Gross, R., Han, E.-H. S., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, J. (1999). Partitioning-Based Clustering for Web Document Categorization. *Journal of Decision Support Systems*, 27(3):329–341.
- Dai, W., Xue, G.-R., Yang, Q., and Yu, Y. (2007). Co-clustering Based Classification for Out-of-domain Documents. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 210–219.
- Mukherjee, A. and Liu, B. (2010). Improving Gender Classification of Blog Authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–217.
- Neubig, G. and Duh, K. (2013). How Much is Said in a Tweet? A Multilingual, Information-Theoretic Perspective. In *AAAI Spring Symposium on Analyzing Microtext*, pages 32–39.
- Nguyen, D., Smith, N. A., and Rosé, C. P. (2011). Author Age Prediction from Text Using Linear Regression. In *Proceeding of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH'11)*, pages 115–123.
- Noll, M. G. and Meinel, C. (2008). Exploring Social Annotations for Web Document Classification. In *Proceedings of ACM Symposium on Applied Computing (SAC)*, pages 2315–2320.
- Orimaye, S. O., Alhashmi, S. M., and Siew, E.-G. (2012). Natural Language Opinion Search on Blogs. In *Proceedings of 12th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pages 372–385.
- Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 17–23.
- Sasa, P., Miles, O., and Lavrenko, V. (2010). The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26.
- Schenker, A. (2003). *Graph-theoretic Techniques for Web Content Mining*. Ph.D. thesis, University of South Florida, Tampa, FL, USA. AAI3182715.
- Uryupina, O., Plank, B., Severyn, A., Rotondi, A., and Moschitti, A. (2014). Sentube: A corpus for sentiment analysis on youtube social media. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 4244–4249.
- Wiebe, J. and Theresa Wilson, C. C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.
- Xiao, M., Zhao, F., and Guo, Y. (2013). Learning Latent Word Representations for Domain Adaptation using Supervised Word Clustering. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 152–162.
- Zhang, Z., Wu, G., and Lan, M. (2015). ECNU: Multi-level Sentiment Analysis on Twitter Using Traditional Linguistic Features and Word Embedding Features. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 561–567.

A Fun and Engaging Interface for Crowdsourcing Named Entities

Kara Greenfield¹, Kelsey Chan², Joseph P. Campbell¹

¹MIT Lincoln Laboratory, 244 Wood St Lexington MA, USA

²MIT, 77 Massachusetts Avenue, Cambridge MA, USA

kara.greenfield@ll.mit.edu, kelseyc@mit.edu, jpc@ll.mit.edu

Abstract

There are many current problems in natural language processing that are best solved by training algorithms on an annotated in-language, in-domain corpus. The more representative the training corpus is of the test data, the better the algorithm will perform, but also the less likely it is that such a corpus has already been annotated. Annotating corpora for natural language processing tasks is typically a time consuming and expensive process. In this paper, we provide a case study in using crowd sourcing to curate an in-domain corpus for named entity recognition, a common problem in natural language processing. In particular, we present our use of fun, engaging user interfaces as a way to entice workers to partake in our crowd sourcing task while avoiding inflating our payments in a way that would attract more mercenary workers than conscientious ones. Additionally, we provide a survey of alternate interfaces for collecting annotations of named entities and compare our approach to those systems.

Keywords: Mechanical Turk, crowd sourcing, named entity recognition, named entity annotation, natural language processing

1. Introduction

Annotated linguistic corpora are a key resource in developing natural language processing algorithms. Many of these algorithms require that their annotated training data is in the same domain as the test data in order to achieve maximal system accuracy. Crowdsourcing platforms such as Amazon’s Mechanical Turk have been shown to be an effective way to quickly and economically gather annotations on text corpora for a variety of annotation tasks. While annotators who have been trained as professional linguists are able to annotate accurately and consistently from dense annotation guidelines, the amateur annotators who serve as workers on crowdsourcing platforms are not similarly motivated to create the best annotations possible. Financial incentives are the most common motivator used with crowdsourcing workers, but it can be beneficial to include alternative incentives as well, such as making the annotation task enjoyable.

2. Named Entity Recognition

Named Entity Recognition (NER) is the subtask of information extraction and consists of automatically extracting named mentions of entities (as opposed to nominal or pronominal mentions) from natural language text.

*This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

The ontology of which types of named entities are to be extracted varies according to application domain. Common ontology sets include Person, Organization, Location; Person, Organization, Location, Date; and Person, Organization, Geopolitical Entity. There have also been several NER systems developed for more specialized ontologies, such as in the medical domain. There are currently several state-of-the-art named entity extractors; however, due to the limited pool of annotated data available, these models are commonly limited to training on formal domains, such as news articles and scientific texts (Finin, et al., 2010); (Nadeau & Sekine, 2007). It is well known that the domain of the training data, which includes both textual genre (journalistic, scientific, informal, etc.) and topic (politics, arts, medicine, etc.) impacts the performance of the system on test data from other domains. For example, Poibeau and Kosseim (2001) showed that some systems yielding F-scores of more than 0.85 on newspaper articles experienced a drop in performance of up to 50% when tested on more informal texts like manual transcriptions of phone conversations and technical emails. Consequently, there is a need for in-language, in-domain annotated corpora with which to train current state-of-the-art NER systems.

3. Traditional User Interfaces for NER Annotation

Most traditional user interfaces for collecting NER annotations allow the annotator to read through the passage once, annotating entity mentions of all classes within the ontology as a single task. Two of the most commonly used off-line annotation tools for collecting NER annotation are the BRAT Rapid Annotation Tool shown in Figure 1 (Stenetorp, et al., 2012) and Callisto (MITRE, 2013). These tools allow the annotator to select a segment of text and then select the appropriate annotation label for that segment. This allows for the annotator to annotate multiple entity types simultaneously, but consequently requires that they mentally keep track of the definitions for those multiple

entity types and go through the process of both selecting the mention and then selecting a label for that mention. Combining the subtasks of annotating mentions of each separate entity type typically saves time for an experienced annotator, who has a good understanding of linguistics in general and the specific definition of the entity classes that they are trying to identify. For novice annotators, such as are likely to participate in a crowd sourcing task, combining tasks can prove to be too difficult, lowering the accuracy of the resulting annotations.

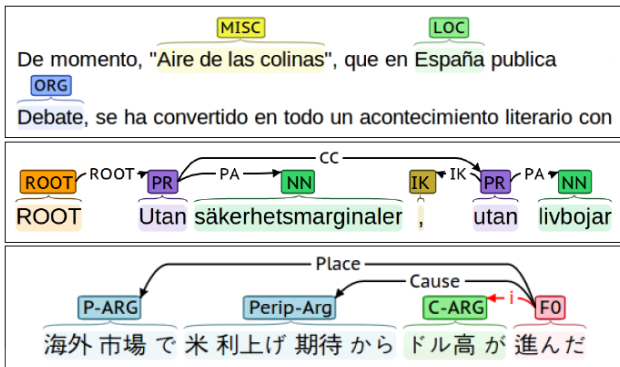


Figure 1 BRAT Rapid Annotation Tool (Stenetorp, et al., 2012)

In addition to off-line annotation tools, there are also several NER annotation interfaces that have been custom-designed for use by crowd sourcing workers on Amazon Mechanical Turk. Some of these are very similar to typical off-line NER annotation tools, requiring the annotator to simultaneously search for entity mentions of all of the types in the ontology. An example of such a system is the Twitter NER Annotation system shown in Figure 2 (Finin, et al., 2010). An interface such as this is relatively easy to create using the built-in requester tools in Mechanical Turk, but forces the annotator to read the passage with one word on each line, limiting the document length that is reasonable to include in a single human intelligence task (HIT). For named entity mentions that consist of only a single token, this interface allows the annotator to indicate as such with only a single mouse click; however an additional click is required for each additional word in the named entity mention.

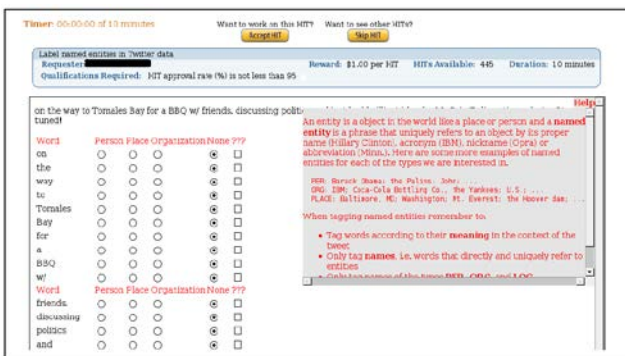


Figure 2 Twitter NER Annotation in Mechanical Turk (Finin, et al., 2010)

An alternative user interface for collecting named entity mention annotations through Mechanical Turk was presented by Lawson et. al. (2010). This was an improvement on previous NER annotation systems in that it included several interface features that were specifically designed to ease the annotation burden on novice annotators, such as Mechanical Turk workers. These features included allowing the user to select spans of text instead of individually clicking on each word and having separate tasks for annotating each type of entity in order to decrease the required mental load. Additionally, this interface had workers annotate both named and nominal entity mentions in an attempt to help workers realize that there is a distinction between named and nominal mentions. This interface can be seen in Figure 3. The usability improvements in this interface were obtained at the cost of needing to create a custom interface for the HITs instead of using one of the default HIT templates. The available templates are not optimal for natural language annotations and the developer cost incurred in creating a custom interface is offset by the resulting increase in annotation quality and decrease in annotation time.

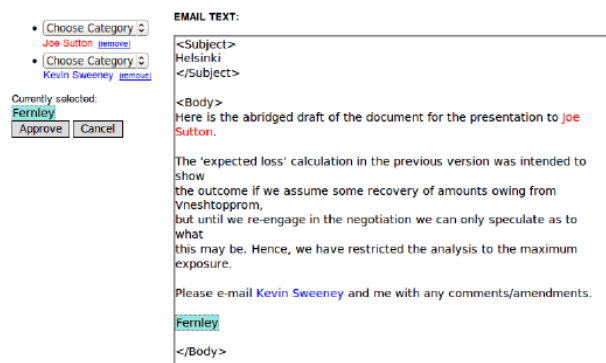


Figure 3 Span-based NER Annotation in Mechanical Turk (Lawson, et al., 2010)

4. MITLL NER Crowdsourcing Annotation System

The MIT Lincoln Laboratory named entity crowd sourcing annotation system maximizes annotation accuracy and efficiency through a combination of 1) a clean user interface that minimizes annotator workload, 2) clear annotation guidelines, and 3) and a methodology for assigning HITs to workers which minimizes low annotation recall.

4.1 User Interface

Our annotation interface built upon the features developed by Lawson et. al. (2010). Our enhancements were primarily focused on minimizing the effort that a worker had to exert in order to annotate a document. By not having workers annotate nominal entity mentions, they were only required to select a text span and click a single button in order to annotate it as a named entity mention. We used color to allow the user to visually see all of the entity mentions that they already annotated and

also the specific text span that they are currently annotating.

We also placed an emphasis on including annotation instructions that were specifically tailored to novice linguistic annotators. Our instructions consisted of a simple definition of a named entity combined with several examples of text spans that were examples of named entities in addition to negative examples. We found that including negative examples in the instructions was particularly beneficial for both increasing annotation accuracy and decreasing the number of workers who emailed us to ask for clarification of the instructions. While detailed instructions are invaluable for assisting the workers, they also require a large amount of screen real estate. We counteracted this by making the bulk of the instructions optionally visible, but always having the simplest form of the instructions (telling the annotator which type of named entity they were supposed to be identifying) visible in large font in a bright color at the top of the screen. Early versions of our experiments didn't have this and resulted in several annotators who otherwise had very high annotation accuracy accidentally annotating the wrong entity type. The system can be seen in Figure 4.

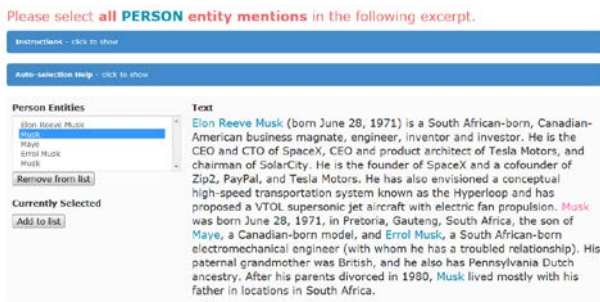


Figure 4 MITLL NER Annotation Interface

4.2 Data Selection and Incentives

Annotator fatigue is a common problem in many annotation scenarios, including crowd sourcing. Failing to counteract this leads to generating annotated corpora that are missing many annotations and consequently can't be utilized as gold standards. This problem occurs even when the annotators are trained linguists, but is compounded in crowd sourced annotation due to the fact that many of the workers are not motivated to care about the quality of the final corpus. Lawson et. al. (2010) addressed this problem by monetarily incentivizing workers based on the number of entities that they annotated. While this methodology did encourage workers to annotate more than just the first few entity mentions in each HIT, it can have the unintended negative consequence of motivating workers to annotate text spans that are not actually entity mentions. The same financial motivations that would lead to a worker not annotating all of the entity mentions in order to annotate more documents when the financial reimbursement is proportional to the number of documents would lead to

those workers annotating an abundance of false positives when the financial reimbursement is proportional to the number of annotations. An additional shortcoming of incentivizing workers based on the number of annotations they return is that the cost of creating the corpus increases by an unpredictable amount.

We primarily chose to address the problem of annotator fatigue by identifying and correcting for it rather than disincentivizing it as Lawson et. al. did (2010). The first way in which we did this (as shown in Figure 5) was to avoid having the same portion of the text occur at the end of the HIT for all of the workers who annotated that HIT. Each document was split into chunks of no more than 500 characters. All excerpts began and ended at sentence breaks so that workers would understand the context of the excerpt. Every HIT contained two excerpts. If an excerpt appeared first out of two in one HIT, it would appear again as the second of two in another HIT. Additionally, we ran all of the documents through our automatic NER system, MITIE (King, n.d.), (Geyer, et al., 2016). We took all of the documents in which MITIE identified entity mentions that weren't annotated by either of the original two workers for that document and presented those sections of text again to a new worker in order to either verify that there was no entity mention or to recover from the low recall of the other workers. Adjudicating automated system output allowed us to benefit from having additional annotations only where they were needed without having to pay to have them on

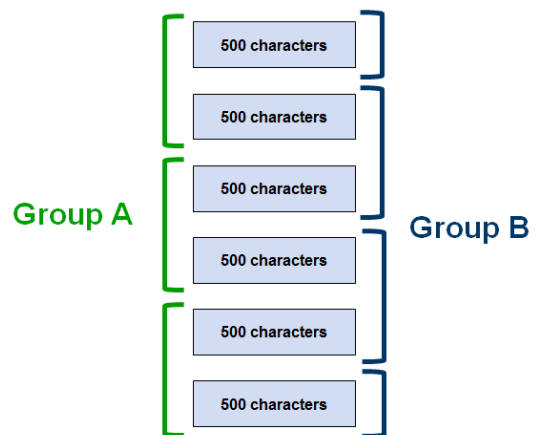


Figure 5 Document partitioning the entire corpus.

We did also appeal to workers' morals and sense of human connection to discourage them from submitting HITs without reading or annotating the text. We accomplished this via a text prompt whenever a user submitted a HIT without any annotations, asking them if they were sure that there weren't any named entity mentions in that HIT.

5. Worker Feedback

We found emails from workers to be an extremely valuable source of feedback on both our interface design

and annotation instructions. While we never explicitly prompted or asked users for feedback, many voluntarily provided it.

One of the greatest benefits that we gained from the pilot runs of our experiments was user feedback on examples in the data where they were unsure of whether or not they should annotate a particular span of text as a named entity mention. In addition to responding to that worker, we used many of those cases as examples in our instructions in the final run of the experiment and correspondingly saw a decrease in such clarification requests which lessened our workload.

Of particular interest was that many of the workers were particularly motivated to maintain their approval rating on Amazon Mechanical Turk. While we didn't say we would reject any HITs or actually reject any HITs, or have any history of ever rejecting HITs on this requestor account, the vast majority of email requests for clarification on the guidelines also informed us that they were diligently trying to complete the HITs accurately and requested that we not reject their HITs if they made a mistake because they were afraid of that negatively affecting their approval rating. There is very likely a positive correlation between a worker being motivated enough to ask for clarification on the guidelines rather than taking their best guess and that worker caring about requestors' opinions of them, so this motivation may not be present in all workers, but it is

Morning :) Just some friendly advice :)
I have done about 140 of your hits. I really like the names ones.
I am guessing your account is a new, based on the # of reviews it has on the workers Turkopticon sight. I also noticed that it seems like your batches are not really being worked as fast as you likely hope, and I wanted to offer some advice on that.
Though I really enjoy your hits (and the interface I must say is really fantastic! Kudos!), the pay does leave something to be desired.

very strong in those who do possess it.

Figure 6 Worker Feedback

We also found that many workers were motivated by the ease of use of the interface, even when they thought that the task warranted a higher financial incentive. Figure 6 shows feedback from one of the workers who completed our HITs. Due to their enjoyment in completing these HITs and the clean interface design, this worker accurately annotated many of our HITs, despite believing that they could obtain a higher hourly rate by completing other HITs. As this worker illuminated, increased financial incentives can serve to decrease the time required to complete a batch of HITs, but with a good

interface design, a slightly lower rate can also yield accurate annotations, just in a slightly longer time frame. While this was the only worker who provided us with feedback on pricing, we received many comments from other workers stating that they found the task enjoyable and especially liked the interface.

6. Conclusions

In this work, we presented a system for gathering named entity recognition annotations via crowd sourcing that builds upon prior work in developing natural language annotation interfaces. We provided a methodology for overcoming the low recall rates that are common among novice annotators. Additionally, we analysed worker feedback to show that having an annotation interface that is easy to use can be a strong incentive for crowd sourcing workers. The primary motivators that we identified other than HIT pricing were maintaining a positive worker rating (which is indirectly a financial incentive) and ease of interface use. In future work, we would like to expand this system to allow for more complicated linguistic annotations, especially those that require annotating multiple disjoint spans of text for a single annotation.

7. Acknowledgements

The authors would like to thank Scott Briere and Nicholas Malyska for their assistance with setting up our crowdsourcing experiments. We would also like to thank all of the Amazon Mechanical Turk workers for completing our HITs and providing feedback to us.

8. References

- Finin, T. et al., 2010. *Annotating Named Entities in Twitter Data with Crowdsourcing*. s.l., s.n., pp. 80-88.
- Geyer, K., Greenfield, K., Mensch, A. & Simek, O., 2016. *Named Entity Recognition in 140 Characters or Less*. s.l., s.n.
- King, D., n.d. *MITLL/MITIE*. [Online] Available at: <https://github.com/mit-nlp/MITIE>
- Lawson, N., Eustice, K., Perkowitz, M. & Yetisgen-Yildiz, M., 2010. *Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk*. s.l., s.n., pp. 71-79.
- MITRE, 2013. *Callisto*. [Online] Available at: <http://mitre.github.io/callisto/index.html>
- Nadeau, D. & Sekine, S., 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1), pp. 3-26.
- Poibeau, T. & Kosseim, L., 2001. *Proper Name Extraction from Non-Journalistic Texts*. s.l., s.n.
- Stenetorp, P. et al., 2012. *BRAT: a Web-based Tool for NLP-Assisted Text Annotation*. s.l., s.n., pp. 102-107.

Novel Incentives for Phrase Detectives

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz and Chris Madge

University of Essex, Language and Computation Group

Abstract

The *Phrase Detectives* Game-With-A-Purpose for anaphoric annotation is a moderately successful example of use of novel incentives to create resources for computational linguistics. In this paper we summarize the *Phrase Detectives* experience in terms of incentives and discuss our future plans to improve such incentives.

1. Introduction

Phrase Detectives (Chamberlain et al., 2008; Poesio et al., 2013) an interactive online **game with a purpose** (von Ahn, 2006) for creating anaphorically annotated corpora through web collaboration, is a moderately successful example of use of novel incentives to create resources for computational linguistics. *Phrase Detectives* has been live since December 2008, collecting almost 3 million judgments on the anaphoric expressions in texts in two languages (English and Italian) from over 40,000 players, resulting in a corpus of over 500 documents and over 300,000 tokens. In this paper we briefly discuss the incentives provided by *Phrase Detectives*, assess their contribution, and discuss future work to address some of the current shortcomings. For further discussion of the incentive structure in *Phrase Detectives* and a more detailed evaluation, see (Chamberlain et al., 2009; Chamberlain et al., 2012; Chamberlain, 2016)

2. A Brief Description of the Game

Phrase Detectives is a single-player GWAP developed to collect data about English (and subsequently Italian) anaphoric reference (Poesio et al., 2013) The game architecture is articulated around a number of tasks and uses scoring, progression and a variety of other mechanisms to make the activity enjoyable. The game design is based on a detective theme, relating to the how the player must search through the text for a suitable annotation (Chamberlain et al., 2008).

The players have to carry out two different tasks. Initially text is presented in Annotation Mode (called *Name the Culprit* in the game - see Figure 1). This is a straightforward annotation mode where the player makes an annotation decision about a highlighted **markable** (section of text). (The annotation scheme used in *Phrase Detectives* is a simplified version of the anaphoric annotation scheme used in the ARRAU corpus (Poesio and Artstein, 2008).)

If different players enter different interpretations for a markable then each interpretation is presented to more players in Validation Mode (called *Detectives Conference* in the game). The players in Validation Mode have to agree or disagree with the interpretation.

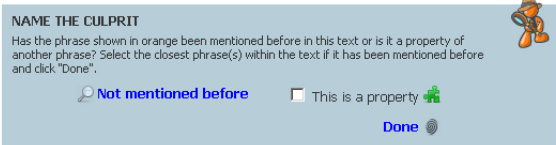
Players are trained with texts from a gold standard. Players always receive a training text when they first start the game. Once the player has completed all of the training tasks they are given a rating (the percentage of correct decisions out of the total number of training tasks). If the rating is above a certain threshold (currently 50%) the player progresses on

Rhinogradentia (Wikipedia)

Rhinogradentia (also known as snouters or Rhinogrades or Nasobarnes) is a fictitious mammal order documented by the equally fictitious German naturalist Harald Stumpke. The order's most remarkable characteristic was the Nasorium, an organ derived from the ancestral species's nose, which had variously evolved to fulfill every conceivable function.

Both the animals and the scientist were allegedly creations of Gerolf Steiner, a zoology professor at the University of Karlsruhe. A mock taxidermy of a certain Snouter can be seen at the Musee zoologique in Strasbourg.

The order's remarkable variety was the natural outcome of evolution acting over millions of years in the isolated Hi-yi-yi islands in the Pacific Ocean.



- Comment on this phrase
- Skip this one
- Skip - closest phrase can't be selected
- Skip - closest phrase is no longer visible
- Skip - error in the text

Figure 1: Detail of a task presented in Annotation Mode.

to annotating real documents, otherwise they are asked to do a training document again. The rating is recorded with every future annotation that the player makes as the rating is likely to change over time. The scoring system is designed to reward effort and motivate high quality decisions by awarding points for retrospective collaboration. A mixture of incentives, from the personal (scoring, levels) to the social (competing with other players) to the financial (small prizes) are employed.

The goal of the game was not just to annotate large amounts of text, but also to collect a large number of judgments about each linguistic expression. This led to the deployment of a variety of mechanisms for quality control which try to reduce the amount of unusable data beyond those created by malicious users, from the level mechanism itself to validation to a number of tools for analysing the behavior of players.

A Facebook version of *Phrase Detectives*,¹ launched in February 2011, makes full use of socially motivating factors inherent in the Facebook platform (Chamberlain et al., 2012). For instance, any of the player's friends who are playing the game form the player's team, which is visible in the left hand menu. Whenever a player's decision agrees with a team member they score additional points. The most

¹<http://apps.facebook.com/phrasedetatives>

interesting finding from this work is that although fewer players play it, the quality and quantity of their work is significantly superior to that of the players of the original game; more in general, knowing the identity of the player leads to much better quality (Chamberlain, 2016).

Phrase Detectives is one of the most successful GWAPs for computational linguistics. Started in December 2008, it is still being played. As of April 2016, over 40,000 players have registered; of these, 4,000 passed the training phase—around 1,000 of which on *Facebook Phrase Detectives*. Over 2.3 million annotation judgments have been collected and 466,000 validations. 549 documents have been completely annotated for a total of around 330,000 words (the complete corpus will be of 1.2 million words). These annotations are being turned into a publically available corpus (Chamberlain et al., 2016).

3. Incentives in Phrase Detectives

The primary incentives in a GWAP for collective resource creation are enjoyment and scientific interest, but we experimented with a number of other types incentives as well. We discuss each in turn.

3.1. Enjoyment

The primary motivation for someone to use *Phrase Detectives* is supposed to be enjoyment: having fun while playing the game. The game was thus designed to incorporate several mechanisms that are meant to make a game fun (Koster, 2005). One of the simplest such mechanisms is **scoring**: by getting a score the player gains a sense of achievement. A second common method to entertain players is to have them experience a **progression through the game**, whether by learning new types of tasks, becoming more proficient at current tasks, or gaining recognition for their effort (see below). A common form of progression is by assigning the player a named **level**, starting from novice and going up to expert (Koster, 2005; von Ahn et al., 2006). (Although we will not discuss quality control here, the level mechanism also provides one form of quality control.) Last but not least, great care was taken in **choosing texts** to annotate that players would find interesting, helped in this by the decision to concentrate on text genres that are under-used in computational linguistics, in particular fiction. We also included a number of documents from Wikipedia, but all chosen for their quirkiness.

3.2. Design

When designing any interface it is essential to know your target audience. Individual, social and socio-technical factors will all determine how successful the interface is at engaging users and what type of data will be contributed. We believe that a key part of the success of *Phrase Detectives* is due to the attractive design of its interface. Game interfaces should be graphically rich, although not at the expense of usability, and aimed at engaging a specific audience (i.e., a game aimed at children may include more cartoon or stylised imagery in brighter colours than a game aimed at adults). Interfaces should also provide a consistent metaphor and work flow. *Phrase Detectives* used a detective metaphor, with buttons stylised with a cartoon detective

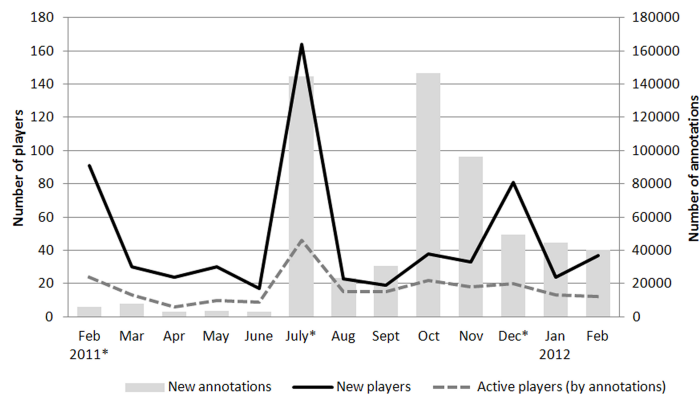


Figure 2: Chart showing the effect of prizes on the workload of *Phrase Detectives* players.

character and site text written as if the player was a detective solving cases. The tasks should be integrated in such a way that task completion, user evaluation and work flow form a seamless experience.

3.3. Contributing to Science

An important incentive for players of GWAPs is the opportunity to participate in a project producing something of relevance to a (scientific) community. This type of incentive did play a role in attracting players to *Phrase Detectives* and retaining them: many of the players of the game are computational linguists who heard about the game through presentations and lectures, or thanks to the mention of *Phrase Detectives* in computational linguistics blogs with a substantial following such as those by Mark Liberman² or Bob Carpenter.

3.4. Prizes

Offering substantial direct payment to the players would defeat the purpose of using GWAPs to reduce the cost of generating resources. But a very low-cost reward structure can be built into online games through the mechanism of **prizes**. In *Phrase Detectives* a variety of prizes in the form of Amazon vouchers for a maximum value of £50 have often been offered. Prizes for high scoring players will motivate hard working or high quality players but the prize soon becomes unattainable for the majority of other players. We also offered therefore lottery style financial prizes, whose winner is randomly selected. In this way the hardest-working players are more likely to win, but the players who only do a little work are still motivated. These prizes have proven extremely effective. Figure 2 shows the effect of prizes on Facebook *Phrase Detectives*. Months where there was active promotion of the site via prizes (February, July and December 2011) show substantial increases in new players, annotations, and active players.

3.5. Social Incentives

A different sort of social incentive is provided by the scoring mechanism. **Public leaderboards** reward players by

²<http://languagelog.ldc.upenn.edu/nll/?p=2050>

improving their standing amongst their peers (in this case their fellow players). Using leaderboards and assigning levels for points has been proven to be an effective motivator, with players often using these as targets (von Ahn and Dabish, 2008). An interesting phenomenon has been reported with these reward mechanisms, namely that players gravitate towards the cutoff points (i.e. they keep playing to reach a level or high score before stopping) (von Ahn et al., 2006).

Both types of social incentives can be made even more effective when the game is **embedded in a social networking platform** like Facebook. In such a setting, the players motivated by the desire to contribute to a communal effort may share their efforts with their friends, whereas those motivated by a competitive spirit can compete against them. This was one of the motivations behind the Facebook version of *Phrase Detectives*.

4. Beyond Phrase Detectives: the DALI Project

The incentives to annotation provided by *Phrase Detectives* could already be defined as having been reasonably successful. The game has motivated a reasonable number of players to annotate a corpus of respectable size. And the corpus already has a significant advantage in comparison with other existing corpora in terms of judgments per markable, with over 20 judgments per markable on average. This said, the ambitions motivating the development of a GWAP are much higher both in terms of number of players (some of von Ahn's games attracted over 100,000 players) and in terms of corpus size (our ambition is to fully annotate over 100 million words). In the soon-to-start DALI project, a collaboration between the University of Essex and LDC funded by ERC, we intend to improve the current incentive structure in a number of ways.

4.1. Making the game more enjoyable

Although many current players enjoy the game, most of those tend to do so because they are interested in the linguistics of anaphora or find the texts quirky, rather than because they find the game enjoyable. Our first objective in DALI will be to develop a new game, or games, which are genuinely enjoyable. Among the ideas we intend to pursue is incorporating in our games a stronger sense of progression, by providing intrinsic rewards to players that achieve a higher status such as the ability to choose more interesting icons for higher status players. We will also develop more attractive ways for players to express their judgments (e.g., clicking on icons associated with discourse entities). We also intend to make smartphones the main platform through which to play the games. While the main motivation for this move is increasing their accessibility, we expect it to make them more enjoyable as well.

4.2. Increased interaction with the computational linguistic community

As mentioned above, a great deal of the success of *Phrase Detectives*, particularly in the beginning, was due to the contribution of the computational linguistics community,

both in popularizing the game through blogs and in actually playing it. We intend to extend the collaboration with the community in collaboration with LDC, both by embedding the game in their future portal for community-created games, and by relying on their expertise in releasing annotated resources.

4.3. Educational Incentives

It can be argued that the most attractive aspect of the current version of *Phrase Detectives* is what it teaches its players about anaphora and its intricacies. This suggests that the game could find a use in teaching language. We intend to test this hypothesis in collaboration with the International Academy at the University of Essex, whose objective is to remedy any language skills shortcomings of future University of Essex students. To this purpose, they offer a variety of language courses that students can take prior to their starting their studies. These courses use a variety of computer-based practice exercises, including games. We recently piloted using *Phrase Detectives* as one of these practice games. We intend to continue and intensify this collaboration.

5. Conclusions

Games with a purpose can serve as a useful alternative for corpus annotation—in fact, as the only viable option when the aim is to create truly large-scale resources (Poesio et al., In press). But in order to realize this potential, sufficient players have to be enrolled through attractive incentives. The first years of the *Phrase Detectives* experience have taught us a lot about what works and what doesn't; we hope to take advantage of these lessons to develop new games that allow us to achieve our objective of creating truly large-scale annotated corpora for computational linguistics.

6. Acknowledgements

The initial funding for *Phrase Detectives* (2007/09) came from UK EPSRC project AnaWiki, EP/F00575X/1. Subsequent funding came from an EPSRC PhD studentship for Jon Chamberlain as well as from the EU project SENSEI. Funding for Chris Madge comes from the IGGI Doctoral Training Centre, funded by EPSRC. DALI will be funded by the European Research Council, ERC.

7. References

- J. Chamberlain, M. Poesio, and U. Kruschwitz. 2008. Phrase Detectives: A Web-based Collaborative Annotation Game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz.
- J. Chamberlain, M. Poesio, and U. Kruschwitz. 2009. A new life for a dead parrot: Incentive structures in the Phrase Detectives game. In *Proceedings of the WWW 2009 Workshop on Web Incentives (WEBCENTIVES'09)*, Madrid.
- J. Chamberlain, U. Kruschwitz, and M. Poesio. 2012. Motivations for participation in socially networked collective intelligence systems. In *Proceedings of CI2012*.
- J. Chamberlain, M. Poesio, and U. Kruschwitz. 2016. Phrase detectives corpus 1.0: Crowdsourced anaphoric coreference. In *Proc. of LREC*, Portoroz, Slovenia.
- J. Chamberlain. 2016. *Harnessing Collective Intelligence on Social Networks*. Ph.D. thesis, University of Essex, School of Computer Science and Electronic Engineering.
- R. Koster. 2005. *A Theory of Fun for Game Design*. Paraglyph.
- M. Poesio and R. Artstein. 2008. Anaphoric annotation in the arrau corpus. In *Proceedings of the sixth International Conference on Language Resources and Evaluation*, Marrakesh, May.
- M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Intelligent Interactive Systems*, 3(1).
- M. Poesio, J. Chamberlain, and U. Kruschwitz. In press. Crowdsourcing. In N. Ide and J. Pustejovsky, editors, *The Handbook of Annotation*. Springer.
- L. von Ahn and L. Dabbish. 2008. Designing games with a purpose. *Communications of the Association for Computing Machinery (ACM)*, 51(8):58–67.
- L. von Ahn, R. Liu, and M. Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of conference on Human Factors in computing systems*, pages 55–64.
- L. von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Evaluation of Anchor Texts for Automated Link Discovery in Semi-structured Web Documents

Na'im Tyson, Jonathan Roberts, Jeff Allen, Matt Lipson

About.com

Sciences, 1500 Broadway, New York, NY, USA

ntyson@about.com, jroberts@about.com, jallen@about.com, mlipson@about.com

Abstract

Using an English noun phrase grammar defined by Hulth (2004a) as a starting point, we created an English noun phrase chunker to extract anchor text candidates identified within web-based articles. These phrases served as candidates for anchor texts linking articles within the About.com network of content sites. Freelance writers—serving as annotators with little to no training outside the domain authority of their respective fields—evaluated articles that received these machine-generated anchor texts using an annotation environment. Unlike other large-scale linguistic annotation projects, where annotators receive an evaluation based on a reference corpus, there was not sufficient time or funding to create a corpus of documents for anchor text comparisons amongst the annotators—thereby complicating the computation of inter-labeler agreement. Instead of using a reference corpus, we assumed that the anchor text generator was another annotator. We then computed the average Cohen's Kappa Coefficient (Landis and Koch, 1977) across all pairings of the anchor text generator and an annotator. Our approach showed a fair agreement level on average (as described in Pustejovsky and Stubbs (2013, p. 131–132)).

Keywords: kappa coefficient, chunking, chunk extraction, link discovery, anchor text

1. Introduction

About.com, also known as *The About Group*, publishes content for various subject domains from topicalized sites across seven major verticals: food, health, home, money, style, tech and travel. The website consists of almost 2 million articles that receive a monthly average of over 200 million visits from visitors primarily in the United States, Western Europe and parts of India. Experts write content in their domain of expertise; with the aid of a content management system, they select snippets of text as anchors to link to other relevant content in their own content website or throughout the entire About.com network.

Given that About.com is a publishing company that makes most of its revenue via advertising, we wish to keep users engaged by pointing them to different parts of the network for as long as possible. Inline links are a critical component of user recirculation—with higher clicks per session—compared to other recirculation methods on the site such as related article listings (at the bottom of an article), trending articles and navigation units around the website.

In our experience though, we found that our experts do not add as many inline links as they could during the process of creating their content. Producing quality links takes a great deal of time, and requires intimate knowledge of the full corpus of About.com content. Usually, experts are not cognizant of related articles written by experts outside of their own topical site. The histogram in Figure 1 demonstrates that the link density of articles (the number of About.com links in a given word count) is typically between 0 and 0.01 prior to the launch of automated link discovery on the site.

The solution was to build a tool that allows experts to select suggested anchor texts in their own articles and choose from the most suitable candidate destinations.

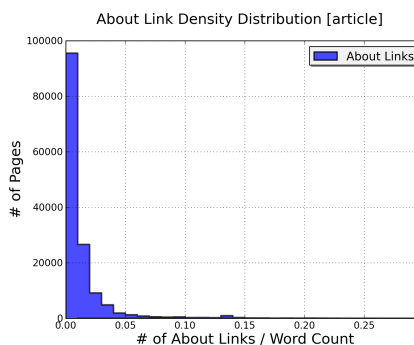


Figure 1: Histogram of link density of articles prior to the launch of automated link discovery.

2. Anchor Text Identification Process

2.1. Previous Approaches

Output from current keyword extraction techniques could serve as a basis for constructing anchor texts within an article given that both anchor texts and keywords encompass small spans of texts. Enhancing keyword extraction with part-of-speech information led to better quality keywords for a database of scientific journal papers (Hulth, 2004b). Other alternatives to linguistically-oriented keyword extraction systems such as KEA (Witten et al., 1999) and TextRank (Mihalcea and Tarau, 2004) might also work as well. The problem with all of these systems is that they tend to generate keyphrases between one and three words in length. In practice, the part-of-speech structures generated in expert-generated anchor texts—exemplified in Figure 2—can differ vastly from smaller noun phrase grammars proposed by Hulth (2004b) and other keyword extraction systems.

Another approach would be to use the existing link knowledge inside About.com to produce anchor and target can-

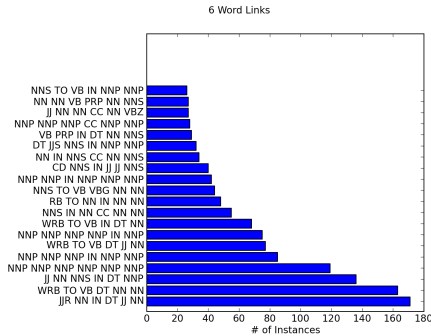


Figure 2: Part-of-speech (POS) histogram of expert-generated anchor texts consisting of six words in full-text articles.

didates. To calculate the strength of an anchor text, we could compute the *target strength* of an anchor text and target document as a ratio of the number of times an anchor points to the target document to the total number of times the anchor text appeared as a link (Erbs et al., 2011). However with so few documents on the site having a sufficient number of links, it would not be worthwhile to implement this technique.

Hence, we devised our own schema for selecting anchor text. All phrases needed to look natural without having any linguistically odd sequences. For example, in the phrase *The Rock and Roll Hall of Fame*, we would not want to use just *Rock and Roll* as a candidate anchor text. An additional requirement was that the method for generating anchor texts had to apply to all subject matter domains referenced on the website; it would be too cumbersome to create different grammar schemes of generating anchor texts for all of About.com’s top-level verticals, and the sites existing within each one.

2.2. Methods for Generating Anchor Texts

2.2.1. Empirically-Driven Approach

Our original implementation had an empirical, linguistically-driven grammar to extract candidate anchor texts—where the grammar sequences originated from existing articles. Figure 2 illustrates an example distribution of grammar sequences. Although these grammar sequences produced longer sequences of anchor texts, they did not consistently identify named entities, and occasionally gave rise to nonsensical anchor texts. Some of these erratic anchor texts appeared due to errors in part of speech tagging.

2.2.2. Chunk Parsing for Anchor Text Generation

An intermediate solution would be to use compound grammatical structures that are less complex than the sequences illustrated earlier, yet general enough to identify potentially complicated grammatical structures. To this end, we used partial sentence parsing, otherwise known as *chunk parsing* (Abney, 1996), to extract phrases from a part-of-speech tagged sentence. Chunk extraction occurred via chunking rules, which are little more than regular expressions of tag sequences, implemented in Python’s Natural Language Toolkit, NLTK (Loper and Bird, 2002). A noun phrase

grammar defined by Hulth (2004a) served as a template for constructing the chunk rules, but it received a great deal of modification and expansion to handle the more complex tag sequences observed on About.com, which included named entities and date/time expressions.

The final anchor text candidates for a document were those having the maximum inverse document frequency across open class words comprising the entire phrase. Candidate destinations for the anchor texts were those having the highest document similarities between the anchor text, and a window of words around it. Because the primary focus of this evaluation was on the quality of the anchor texts, we will not concentrate on the exact method of computing similarity between source and target documents in this paper.

3. Evaluating Anchor Texts

3.1. Quality Assurance Setup

Before deploying automated link discovery throughout About.com, we decided to implement a Quality Assurance (QA) phase to adjust our algorithm for anchor text generation. This QA phase included 13 freelancers, who served as annotators, to verify anchor texts from approximately 86,000 articles chosen from our most highly viewed content on the site.

3.2. Annotation Workflow

In a similar fashion to Huang et al. (2009), where annotators had the opportunity to select link targets, and mark anchors and targets as relevant or irrelevant, our annotators had the following options within a web-based annotation environment: 1) keep an anchor text, 2) modify an anchor text by expanding or contracting it, 3) delete it entirely and 4) modify the link target. Annotators saw a single link target that they could delete, or supply one of their own. Usually annotators for tasks such as these would receive a great deal of training to ensure they could properly and consistently identify possible anchor texts in documents. In these circumstances, though, having few available options for the freelancers to mark up anchor texts and link targets inside the annotation environment made the need for further training somewhat of a burden—especially in light of the schedule to re-publish the documents with their enhanced links. Freelancers received payment on an hourly basis, and did not garner additional wages upon the project’s completion. The hourly incentive obviated the desire to annotate documents in haste. A database connected to the annotation environment tracked annotations across all of the freelancers’ sessions; this gave content managers who managed the final documents the ability to undo certain annotations at some later time if they saw that the revisions were nonessential.

3.3. Evaluating Anchor Texts for Inter-Labeler Agreement

Evaluating the anchor texts in isolation proved to be a difficult task because the complete validation required some consideration of the link target. Assuming that the link target was satisfactory, then the previously mentioned options for altering the anchor text remain the same. If we used the entirety of the anchor text as the unit for evaluation, we fail

to give credit to the generator when there is slight disagreement on the span of an anchor text. Consequently, the evaluation treats the anchor as a sequence of words to measure the relative agreement between the generator and an annotator. If the words in an anchor text remain unchanged, the relative agreement is one. We did not consider anchor texts with deleted link targets since we had no way of knowing why the annotator deleted the target link: the target link may be inappropriate for the anchor text, or the target link may not have fit the context of the article.

3.4. Computing Inter-labeler Agreement

Annotators were not privy to the anchor texts deemed unsuitable for linking by the generator, so there is no way to directly measure when both the generator and the annotator identified anchor texts as negative. As an approximation, each anchor text received a padding of one word before and after the text to estimate words that either the generator or annotator ignored. A caret and a dollar sign denoted the padded token at the beginning and end of phrases, respectively, as illustrated in Example 1.

Symbols a through d in the same example refer to cells in a contingency table, shown in Table 1, for each phrase extracted from a document. The letter 'A' denotes the generator, and 'B' represents an annotator; and the 'positive' label identifies an agreement between both annotators. Our assumption was that the annotator represented ground truth. The cell marked 'a' is the relative number of words where the generator and annotator agreed on the anchor text; we can consider this as the relative number of *true positive* words between the automatically generated anchor and the annotator's selection. Cell 'b' is a relative number of words that the generator suggested as anchor text, but the annotator modified or deleted it. These are the words that the generator falsely identified as anchor text and the annotator ignored—thereby making these words *false positives*.

When the generator did not select words in the anchor text, but the annotator inserted words, then we measured that relative disagreement in cell 'c', and called them *false negatives*. Padding tokens at the beginning and end of the anchor text selected by the algorithm and annotator—which indicate the outer boundaries of the anchor text—gave us the ability to approximate the number of *true negatives* between the algorithm and annotator for cell 'd'. Table 1 shows the placement of symbols a through d within a two-way contingency table; and Table 2 is an instantiation of Table 1 with the relative number of correct/incorrect words derived from all 11 words presented in Example 1.

Algorithm: ^ quick brown fox jumps over the lazy dog \$
 Annotator: ^ the quick brown fox \$
 d c a a a b b b b b d

Example 1: Example of phrase alignment between the anchor text generator and an annotator.

With a two-way contingency table for each pair of annotators, i.e., for the generator and the human annotator, we computed the *mean average precision*, MAP, and Cohen's Kappa (K) as shown in Equations 1 and 2, respectively.

		B	B
		positive	negative
A	positive	a	b
A	negative	c	d

Table 1: Contingency table for the anchor text generator (A), and a single annotator (B).

		B	B
		positive	negative
A	positive	3/11 = 0.27	5/11 = 0.45
A	negative	1/11 = 0.09	2/11 = 0.18

Table 2: Contingency table computed from relative word agreements from Table 1 for the generator (A) and annotator (B).

$$MAP = \frac{1}{|A|} \sum_{i=1}^{|A|} \frac{1}{m} \sum_{k=1}^m Precision(d_k) \quad (1)$$

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (2)$$

For the MAP calculation, we computed the average precision per document, $Precision(d_k)$, and then averaged across all annotators in the set A .¹ This gave us a MAP score of 0.40.

In Equation 2, $Pr(a)$ is the ratio of agreement between the annotators and the total number of annotations. In short, $Pr(e)$ is a summary metric of the expected agreement for each category label. It first involved calculating the percentage of times that the annotator used a particular label; this was the equivalent of summing across a row or column for a category and dividing by the total number of annotations. Since each annotator worked independently, we took the product across annotators. Finally, there was a summation of the product for each category across all categories because the distribution of the categories is disjoint.² Averaging over the kappa coefficients for the pairs of anchor text generator and annotator should yield a fair to moderate agreement level.

The average Kappa from all of the documents was 0.33, which would only be a *fair level of agreement* between the generator and the annotator. A preliminary chi-squared test on the relative agreements from the generator and the annotators showed that there was a difference between the relative agreements of the generator and annotators; therefore, the generator's choices are not always on par with those of the annotators.

¹Precision is the number of true positions divided by the sum of true and false positives ($tp/tp + fp$).

²See Pustejovsky and Stubbs (2013, p. 133–134) for a detailed example of how to compute both $Pr(a)$ and $Pr(e)$.

4. Discussion

4.1. Incentivizing a Two-tier Approach to Web Page Annotation

In contrast to other Language Resource construction projects, where contributors/annotators have some social aspect to their work, the annotation tasks—as we described them here—are very solitary in nature. A freelancer logs into the annotation environment to validate anchor texts and linked targets, and nothing else. They received little to no feedback from the application or supervisor during the annotation process.

To ensure labeling consistency for the words comprising anchor texts, while promoting a social aspect to the process of annotation, we could divide the QA process into two subprocesses: a test for validation consistency and the actual QA process (as described in Section 3).

In a test of consistency, freelancers or content authors would receive the same set of documents for mark up. These documents would already have links inserted in them using the automated linking process; and this is no different than what we described earlier (but with far fewer documents). The difference here is that we would measure the *mean average relative agreement*, MAR, between each pair of annotators, and exclude the anchor text generator.

Using just the agreements a and d from Table 1 for an anchor text, t , we can compute the *average relative agreement* for a document consisting of each anchor text, t , with Equation 3:

$$\frac{1}{n} \sum_{i=1}^n a_{t_i} + d_{t_i} \quad (3)$$

Dividing the sum of a and d by the sum of a , b , c and d is not necessary since the agreements are relative, as shown in Table 2, and the denominator would have a sum close to one. Finally, we take the mean across the set of all documents, D , to compute the mean average relative agreement between any two annotators:

$$MAR = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{1}{n} \sum_{i=1}^n a_{t_i} + d_{t_i} \quad (4)$$

With a MAR measurement for every pair of annotators, we can use a threshold to compare scores to find bad actors within the group (Neuendorf, 2002). If the MAR regularly falls below a threshold, the annotator would not receive an extra incentive to continue with the rest of the annotation of the corpus. Content managers could hold a general meeting among the annotators in order to expose good and bad practices in annotation, and allow the annotators to meet face-to-face.

4.2. Improvements to Anchor Text Selection

Thus far, we focused on honing the QA process to increase annotator consistency and compensation. To raise the level of agreement between the human annotators and the anchor

text generator, there are a few options we can explore for enhancing the generator.

First, we could offer more parses of a sentence given the noun phrase grammar we constructed. NLTK returns a single parse of the sentence that matches the first rule within our noun phrase grammar. We could submit a pull request to the NLTK GitHub Project that fixes this issue. This requires a long-term commitment that we have to schedule into a future software release.

An alternative to this massive software enhancement would be to build a probabilistic noun phrase grammar in NLTK. Such an effort entails computing probabilities of noun phrase constructions from the existing anchor texts that already exist on the site. If there was not a sufficient number of examples for each noun phrase construction, we could turn to the anchor texts used as a result of the annotations, along with smoothed probabilities to accommodate those constructions where there were still not enough examples within the corpus.

5. Conclusions

We presented a novel framework for evaluating anchor texts generated by an automated link discovery system for the purpose of computing inter-labeler agreement. This evaluation scheme yielded only a fair level of agreement between the anchor text generator and the annotators we employed during the quality assurance phase of the automated link discovery system. With a reference corpus and better incentives offered to the annotators, accompanied by enhancements to the anchor text generation process, we hope to achieve a higher level of agreement in the foreseeable future.

6. Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments.

7. Bibliographical References

- Abney, S. (1996). Tagging and partial parsing. In Ken Church, et al., editors, *Corpus-Based Methods in Language and Speech*. Kluwer Academic Publishers, Dordrecht.
- Erbs, N., Zesch, T., and Gurevych, I. (2011). Link discovery: A comprehensive analysis. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 83–86, Sept.
- Huang, W. C., Trotman, A., and Geva, S. (2009). The importance of manual assessment in link discovery. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09*, pages 698–699, New York, NY, USA. ACM.
- Hulth, A. (2004a). *Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction*. Ph.D. thesis, Department of Computer and Systems Sciences, Stockholm University.
- Hulth, A. (2004b). Enhancing linguistically oriented automatic keyword extraction. In *Proceedings of the Human Language Technology Conference*. North American

- Chapter of the Association for Computational Linguistics.
- Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Philadelphia: Association for Computational Linguistics.
- Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing Order into Texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.
- Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, California: Sage Publications.
- Pustejovsky, J. and Stubbs, A. (2013). *Natural Language Annotation for Machine Learning*. O’Reilly Media, Inc.
- Witten, I., Paynter, G., Frank, E., Gutwin, C., and Nevill-Manning, C. (1999). Kea: Practical automatic keyphrase extraction. In *International Workshop on Description Logics*, pages 254–256.

Unconventional Approaches to Gathering and Sharing Resources for Spoken Dialog Research

Maxine Eskenazi, Sungjin Lee, Tiancheng Zhao, Ting Yao Hu, Alan W Black

Language Technologies Institute – Carnegie Mellon University

5000 Forbes Ave Pittsburgh PA 15213 USA

E-mail: max,awb @cs.cmu.edu, tianchez, tingyaoh@andrew.cmu.edu, junion@yahoo-inc.com

Abstract

The DialRC and DialPort projects have employed unconventional approaches to data gathering and resource sharing. The projects started sharing by distributing the speech, transcription and logfile data gathered by the Let's Go system. That system has responded to over 220,000 calls from real users of the Allegheny County Port Authority. The Let's Go platform proved to be a very successful way to run studies, with a dataflow of about 1300 dialogs per month. Thus, DialRC built a research platform that was used by other researchers, enabling them to run studies with the Let's Go real users. Challenges were also run on this platform. Finally DialPort follows in the footsteps of DialRC by creating a spoken dialog portal with real users that other dialog systems can be connected to. This paper examines the impact that these activities have had on the spoken dialog research community.

Keywords: spoken dialog, portal, data sharing

1. Introduction

Over the past ten years, the Dialog Research Center has taken unconventional approaches to gathering and sharing resources. The Center has focused on providing the means for researchers from other sites to share data and run studies. Data gathering is a novel approach with speech from real spoken dialog system users being logged and distributed. The novel approach to running studies centers around opening up access to the Let's Go platform by both distributing its software and inviting researchers from outside Carnegie Mellon to run studies on it.

Data gathering and sharing began with the Let's Go project (Raux et al 2006). Working with the Port Authority of Allegheny County (PAT), the team created a telephone-based spoken dialog system that answers the phone for PAT callers in the evenings and on weekends. It gave bus schedule information when humans were not working. The system went live on March 5, 2005 and is still functioning. It has been "live" every day except one (machine room flooding) and has expanded to 24/7 availability. In the fall and winter of 2010, the coverage was expanded from 10 bus routes in the East End of the city to 60 routes that pass through the East End during some part of their trip. In the summer of 2014, it became directly accessible via a phone number that has been advertised on the buses. The system has logged over 220,000 dialogs over its 11 years of existence.

The dialog with Let's Go is relatively simple, but has been found over the years to still be complex enough to study interesting research issues. The users provide

a time, a departure stop, a destination stop and, optionally, a bus route and the system provides the times and the number of the appropriate bus. The system deals with a wide variance on how bus stops are described by users and it must respond to real callers and the consequences of real telephone background noise from crying babies, loud TVs, and traffic noise.

The team that built Let's Go carried out many studies on it. It was later made available to others in the community. In this way, Lets Go has provided: real user data; the system software; and a platform on which to run studies.

2. The Dialog Research Center Activities

The Dialog Research Center (DialRC) was formed with a grant from the National Science Foundation. The goal was to make the Let's Go products widely available to the research community. The following sections give more detail on those products

2.1 Speech and logfile data from the live Let's Go system

The data consists of speech files, both of the whole dialog and also of each user turn. They are accompanied by the system logfiles for the corresponding dialogs. An interface relates the two, where a summarized logfile is viewed and each

corresponding turn can be heard. The data is available in two forms. There is a small test set that is directly available on the web. The over-700GB dataset is available by exchange of hard drives. The latter has so far been distributed to 17 groups throughout the world. Some of the data has been labeled. About 6 months of calls have been labeled by an expert. One year of calls has been labeled through crowdsourcing (Parent and Eskenazi, 2010). Other data was labeled for use in the Challenges described below.

2.2 MyBus software

A simplified version of the Let's Go system was created to be used to teach students about spoken dialog systems. MyBus uses that Olympus spoken dialog architecture (Bohus et al 2007) and was first distributed at a tutorial presented at HLT 2008 (Raux et al 2008). The software is downloadable, has a wiki for discussion and members of the DialRC group field questions from its users. MyBus could also be integrated in a course on spoken dialog. Since the software provides a good introduction to Let's Go, some researchers have used it to prepare for the creation of a full blown system that they later ran for a study on Let's Go live.

2.3 The research platform

DialRC made the Let's Go system available to other researchers. They prepared their version of the system and the DialRC team tested its robustness. When the system passed, being at least as robust as the live system at the time, it was allowed to "go live". The platform was used for tests of real vs paid users (Ai et al 2007) and of lexical entrainment (Stoyanchev 2009, Stoyanchev and Stent 2009), for example. It took time for researchers to accept the vision of a commonly-shared platform. When it was finally accepted as a new paradigm, it became the source of the Spoken Dialog Challenge (Black and Eskenazi 2009).

2.4 The Spoken Dialog Challenge

The Spoken Dialog Challenge (SDC) (Black et al 2010, Black et al 2011) was designed to bring together spoken dialog system researchers on a common task. Since comparisons of dialog and evaluation techniques are hard to carry out between different systems, different domains and different user populations, the goal was to offer one domain and user population that allows more common bases of comparison. The goal was also to provide large quantities of real user data as the basis of

comparison. The Challenge was not seen as a competition, but rather a comparison of methodologies. Thus each group built a Let's Go system of its own and ran it live on the Let's Go phone number.

The Spoken Dialog Challenge 2010 was divided into three stages: development, control testing, and live testing. For development, the full source code for the system was released as well as the data (the text logs as a download, and the text plus audio as a disk mailing service). 10 groups received the data. Participating groups could use their own dialog architecture if they desired.

Even groups who didn't build on the existing Let's Go source code found the Let's Go language models, grammars, etc. very useful. In the end, four groups produced working systems for the control tests (two universities in the UK, one industry research lab in the US, and CMU's base system). The control tests used spoken dialog experts to call each system with a given scenario. Although the completion rate was higher than for live tests, most of the callers were not from Pittsburgh, and many were non-native speakers (or spoke non-US dialects of English).

The initial results of the systems (on the control tests) were presented at a well-attended special session at SLT2010 in Berkeley, CA in December 2012, while the final live test results (after hand-labeling all of the dialogs) were presented at SIGDIAL2011 in Portland, OR in June 2011. Although WER rate generally correlates with task completion, there were different system orderings for task completion depending on control or live tests. This again highlights the observation that optimization for lab test subjects may not reflect the outcome with real users.

The second Challenge, SDC2011, had a total of 4 participating systems, though these came from only two institutions. This allowed closer comparisons of specific system differences, but did not offer the breadth of systems that participated in the first year.

The clear theme of the SDC2011 participating systems was **dialog state-based** techniques. Although some general changes were necessary in the default system from the first year (due to schedule changes, and increased route coverage), the default system was fundamentally the same as SDC2010 so some cross-year comparison was possible. Since both teams had had experience in producing robust systems, control tests were not carried out. The live tests took place from December 2011 to February 2012. The two groups (four systems) taking part in SDC2011 submitted result papers to SLT2012 (Miami, FL, December 2012).

The rise of the interest in dialog state during this

Challenge gave rise to a new type of Challenge, the Dialog State Challenge.

2.5 The Dialog State Challenge

The first and second Dialog State Tracking Challenges (DSTC) (<http://research.microsoft.com/en-us/events/dstc/>) is a follow on from Spoken Dialog Challenges 2010 and 2011. A number of researchers in the domain wanted a means to have better comparisons and to accurately estimating a user's goal in a spoken dialog system. Having a common task and a common large dataset answered this need. The results of the Challenge were presented at SIGDIAL 2013. They used the Let's Go data and DialRC provided support for the data and its annotation.

2.6 The impact of DialRC on the spoken dialog research community

Since the goal of DialRC was to serve the community, its success can be measured by how much its products were used by the spoken dialog community.

In order to determine impact, a targeted search of the literature in spoken dialog was conducted. This reflects both how aware the community is of the DialRC approaches (gaining mention in a paper) and whether they have actually used the products (paper results being based on their use). As mentioned above, those products are:

- the distributed speech, labels and system log data,
- the MyBus/Let's Go system,
- studies run on the Let's Go platform,
- and participation in the Spoken Dialog Challenge and/or the Dialog State Challenge.

The assessment below, carried out in 2013, (marking the end of the DialRC funding) refers to papers found using keywords such as "Let's Go", "DialRC", names (authors of the Let's Go papers), and "dialog challenge". For publications between 2008 and 2012, a total of 216 references (non-CMU publications) were found. Figure 1 shows the total number of publications by year. There is a steady increase over the years.

Figure 1 indicates that there is significant awareness of the DialRC products within the research community. To determine whether the products were being adopted and actually *used* for publications, the papers were read by the DialRC team. Figure 2 breaks the data in Figure 1 down into two parts: the references that simply *mention* DialRC products and those that actually *use* them. We see that mention of

DialRC started out strong in 2008 and increased from 2009 to 2012. In 2008 there were few authors who actually *used* DialRC products, but this increased in the following years. It is interesting to note that some of the authors, who only mentioned the products one year, went on after that to actually use them. In 2012 the number of product users seemed to have leveled somewhat, while the total number of publications increased.

More detailed examination of the publications from 2009 and 2012 reveals a wide range of topics. The dialog research community has varying and changing interests (e.g. from simulated users to multiparty dialogs). The references to DialRC covered *eleven different topics*, as seen in Figure 3. Although a large portion (116) were about system architecture, we note that discourse (19), user behavior (20) and evaluation (32) were also well-represented.

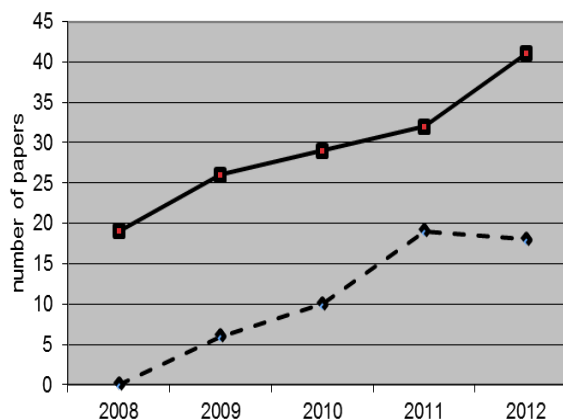


Figure 1. References by year of appearance

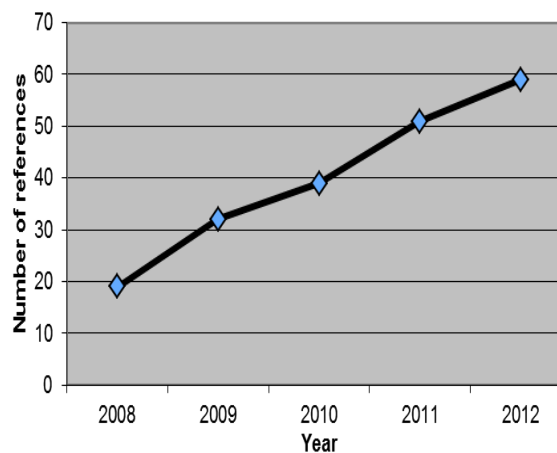


Figure 2. Total references (dashed line) that use the DialRC products and total references that mention (solid line) the products by year.

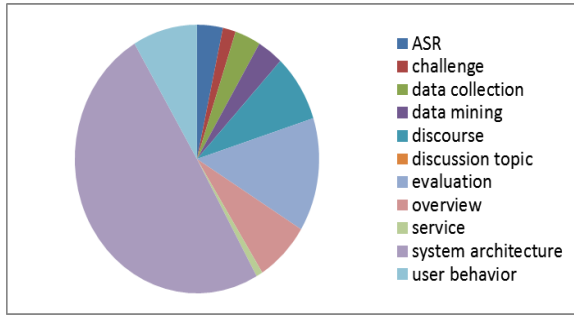


Figure 3. DialRC topics in spoken dialog research – from the 2009-2012 literature

In the long term, DialRC use should result in substantial contributions, such as journal papers, theses and book chapters. Although we see that conference (112), symposium (5) and workshop (44) papers have indeed been the most prevalent (two of the conference papers are main keynotes at Interspeech, Steve Young 2010, Julia Hirschberg, 2011), we also note a reasonable number of references in journal papers (40) and book chapters (12). And, interestingly, there are many theses (18). Nine of the theses actually used the system or the data as an integral part of the thesis work.

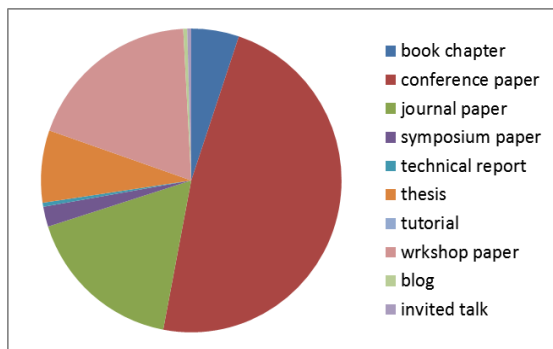


Figure 4. DialRC and types of publications

2.7 DialRC products beyond 2012

Even after the end of DialRC funding, researchers have continued to use its products. For the period from 2013-2015, Figure 5 shows the persistence of the influence of DialRC. The results shown above for 2012 are included in this Figure for reference. There are a total of 92 papers that mention DialRC products over this three year period. There are 31 that actually use the products. We see that there is a gradual

decline in the number of papers mentioning the products and a decline and then steady actual use of the products. This is partially due to the Dialog State Challenges and to the distribution of a Let's Go user simulation. It can also be attributed to a database that was built in Germany using the Let's Go material.

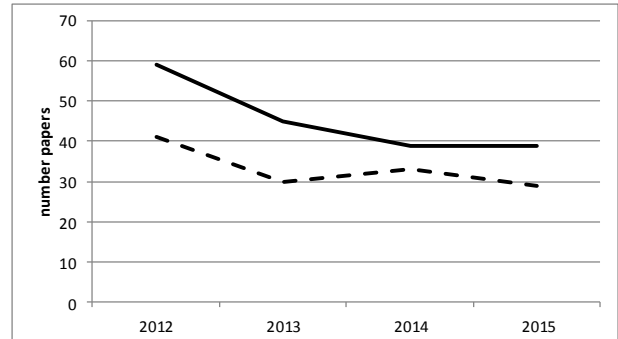


Figure 5. Use of the products after the end of the DialRC grant– solid line is all papers mentioning the DialRC products, dashed line is papers using the products.

3. A Spoken Dialog Portal

We see from the large number of publications that the products distributed by DialRC fulfilled a need in the spoken dialog research community. But the community and its research needs evolve. In the years since DialRC and the study above, there have been more spoken dialog systems, spurred on by the advent of SIRI and other personal assistants.

While it is relatively easy for industrial systems to get real users, academia has more difficulties. And the academic systems are very diverse, going from web-based to phone-based to app-based to robot-based. Given the DialRC team's long experience with real users, it was natural to evolve from getting users for one system to getting users for the research community. The concept of a portal began to take shape. The idea was to have a single user-directed interface that would link all possible academic systems. In this way, the cost of attracting users could diminish and the added-value to the user of finding many different types of assistants would be attractive. The DialPort project was born and Skylar came into existence.

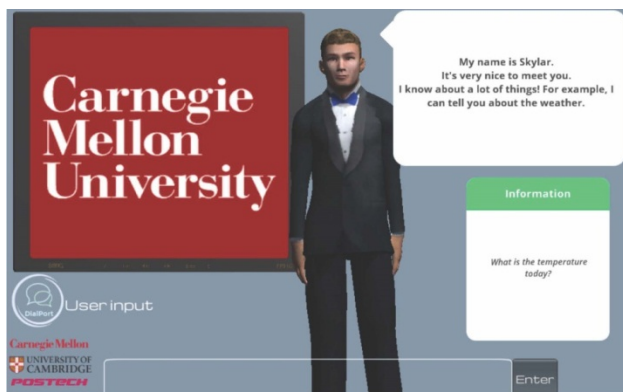


Figure 6. Skylar, the butler, is a portal where real users can speak to academic dialog systems

Skylar (Figure 6.) is the avatar of the portal. The user meets Skylar upon entering the portal and it is Skylar's job to determine what the user wants and to convince them to try all of the connected spoken dialog systems. Since keeping a real user engaged involves providing a large quantity of varied information, Skylar knows about the weather and hotels and restaurants. It is also endowed with a chatbot. But every few turns, it tells a user that it has a friend who can give them information about, for example, hotels in San Francisco (the Cambridge University system). It encourages the user to ask about that information. When the user does ask about another system, another avatar appears (chosen by the system developers amongst the characters available in the Unity software) to talk to the user. The transfer appears seamless to the user and the control of the dialog goes back to Skylar when the other dialog is finished.

Characterized as a butler, Skylar's movements are coded to resemble those of television butlers so that its positions, for example, when it is listening, are easy to interpret by any user. At present Skylar is linked to the Cambridge system. In the summer of 2016, it will be linked to Let's Go. In the fall of the same year, it will be linked to two more systems. Just as the first connection was an interesting challenge, the Cambridge system and the Let's Go system are phone-based. Later connections involving apps and/or robots should provide further interesting challenges.

4. Conclusion

The DialRC products and the DialPort activities are creating novel research opportunities. The platform has given the DialRC team a large real user base and has afforded many studies that could not have been

carried out without this quantity of data. Annotating the data gave the team experience in crowdsourcing. Finally the portal is giving the team experience in interfacing systems of very different natures.

5. Acknowledgements

DialRC and DialPort are funded, respectively, by NSF grants CNS-0855058 and CNS-1512973. The opinions expressed in this paper do not necessarily reflect those of NSF.

6. Bibliographical References

- Ai, H.; Raux, A.; Bohus, D.; Exkenazi, M.; and Litman, D. 2007. Comparing Spoken Dialog Corpora Collected with Recruited Subjects versus Real Users. *In Proc. 8th SIGDial Workshop on Discourse and Dialogue*, Antwerp, Belgium.
- Black, AW, Burger, S, Conkie, A, Hastie, H, Keizer, S, Lemon, O, Merigaud, N, Parent, G, Schubiner, G, Thomson, B, Williams, JD, Yu, K, Young, S, and Eskenazi, M, 2011, *Spoken Dialog Challenge 2010: Comparison of Live and Control Test Results*, Proc. SIGDIAL 2011, Portland.
- Black, A., Burger, S., Langner, B., Parent, G., Eskenazi, M. (2010). Spoken dialog challenge 2010. Proceedings of SLT 2010, Berkeley, CA,
- Black, A.W. and Eskenazi, M. (2009). The spoken dialogue challenge. Proc SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL '09), London, UK. pp. 337-340.
- D. Bohus, A. Raux, T. Harris, M. Eskenazi, and A. Rudnicky, 2007, Olympus: an open-source framework for conversational spoken language interface research, HLT-NAACL 2007 workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology, Rochester, NY, USA.
- Eskenazi, M., Black, A. W., Raux, A., Langner, B. (2008). Let's Go Lab: a platform for evaluation of spoken dialog systems with real world users. Proc Interspeech 2008, Brisbane, Australia, p. 219.
- Parent, G. and Eskenazi, M. (2010). Toward better crowdsourced transcription: Transcription of a year of the Let's Go bus information system data. Proceedings of SLT 2010, Berkeley, California.
- Raux, A., Bohus, D., Langner, B., Black, A., Eskenazi, M., 2006, Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience, Proc. Interspeech 2006, Pittsburgh.
- Raux, A., Langner, B., Black, A. and Eskenazi, M. (2008) Building practical spoken dialog systems.

- Proceedings of ACL/HLT 2008 Tutorial, Columbus, Ohio, United States.
- Stoyanchev, S. (2009). Impact of responsive and directive adaptation on local dialog processing. PHD Thesis, State University of New York at Stony Brook, NY, United States.
- Stoyanchev, S. and Stent, A. (2009). Lexical and syntactic priming and their impact in deployed spoken dialog systems, In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers (NAACL-Short '09). Boulder, CO, United States, pp.189-192.

7. Language Resource References

The Let's Go data is distributed by the DialRC.
<https://dialrc.org/>.

Oral Histories: Linguistic Documentation as Social Media

Mark Liberman

Linguistic Data Consortium. University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
myl AT ldc.upenn.edu

Abstract

Oral history recordings were pioneered by anthropologists in the early 20th century, collected by Alan Lomax and by the Federal Writers' Project during the 1930s and 1940s, and popularized by authors like Oscar Lewis and Studs Terkel in the 1950s and 1960s. Inexpensive tape recorders allowed the form to spread in the 1960s and 1970s. Now a new opportunity is provided by the combination of ubiquitous multimedia-capable digital devices, inexpensive mass storage, and universally accessible networking. The potential popularity of oral history-like recordings is demonstrated by the tens of thousands of people who have made recordings for StoryCorps. However, there is still no easy way for a motivated group – a family, an athletic group, a school class, a business, a scholarly discipline a club, a church -- to create and publish a collection of oral histories or similar forms of cultural documentation. But the software required to make and edit such recordings, and to transcribe, index, document, publish, and comment on them, is relatively simple and easy to create. And with the right infrastructure, millions of people around the world would participate, creating linguistic and cultural documentation on an unprecedented scale.

Keywords: language resources, collection, annotation, transcription, distribution.

1. Introduction

Oral history recordings were pioneered by anthropologists not long after the first primitive recording devices became available. John Lomax taught the Federal Writers' Project to use oral-history interview techniques in collecting thousands of first-person life stories of ordinary people during the Great Depression. His son Alan Lomax collected oral history interviews along with his folk-music recordings in the 1930s and 1940s. Books and radio programs by authors like Oscar Lewis and Studs Terkel popularized the form in the 1950s and 1960s. Inexpensive tape recorders allowed the form to spread in the 1960s and 1970s, and enabled sociolinguists and dialectologists to record tens of thousands of hours of interviews, though there was still no easy way to reproduce and share the results.

Now a new opportunity is provided by the combination of ubiquitous multimedia-capable digital devices, inexpensive mass storage, and universally accessible networking. Anyone with a smartphone can make good-quality recordings and upload them to the cloud. And the potential popularity of such recordings is demonstrated by the fact that tens of thousands of people have made (alas mostly inaccessible) recordings for the StoryCorps project, and millions of people have listened to selected and edited samples on NPR.

However, there is still no easy way for a motivated group – a family, an athletic group, a school class, a business, a union, a political campaign, a scholarly discipline, a club, a church -- to create and publish a collection of oral histories or similar forms of cultural documentation. But the software required to make and edit such recordings, and to transcribe, index, document, publish, and comment on them, is relatively simple and (where not already existing) will be easy to create. And with the right infrastructure, millions of people around the

world would participate, creating linguistic and cultural documentation on an unprecedented scale.

1.1. Existing archives

There are hundreds of thousands of hours of existing untranscribed recordings. Many of these have been preserved in digital form, and others are being digitized – the question of how to salvage and preserve this material is an issue for another time. Tens of thousands of hours of sociolinguistic and dialect-survey recordings are now available in digital form, or soon will be.

For nearly all of these recordings, there's a natural constituency of interested people – ordinary citizens as well as linguists, historians or sociologists – who could be enlisted to help with transcription and annotation, if there were an efficient way for them to participate in distributed web-based projects.

In addition, there are many formal oral history projects that have been completed over the years, but are in many cases not easily available due to lack of any easy way to publish them, or because of outdated and inappropriate attitudes towards intellectual property or “human subjects” restrictions.

1.2. New collections

Recently, a colleague was initially interested in gathering oral texts for a lexicographic project in a widely-spoken but under-documented language. I showed her how to use a cell phone or tablet to make recordings. The first person that she interviewed was someone who had participated in an important series of historical events half a century ago.

At the end of an enthusiastic hour-long interview, the subject said “But I have so many more stories to tell, and my friends have so many more!” The interviewer found that first hour of stories so interesting that she decided to spend the summer making similar interviews in

her home country, and to look for ways to publish the results.

This is a common reaction. Most people have interesting stories to tell, if you give them a chance; and given the opportunity, most people like listening to other people's stories. And there are many social groups and events that motivate thematic collections of stories: intellectual or political movements, athletic teams, neighborhoods, school or church groups, professional gatherings, family reunions, and on and on. Whatever the topic or the occasion, there are millions of individuals and thousands of groups who would create accessible archives of oral texts if it were easier for them to do it.

A lot of this is already happening on tumblr, YouTube, Facebook and so on. But that material is generally shorter, less permanent, and/or less organized than the sort of thing envisioned here.

2. What we need

Some of the needed components already exist. Would-be collectors can easily find

- ways to record and edit audio and video
- ways to share audio, video and text online
- ways to organize online discussions

Some things that are not as easily available are

- good tools for transcription and alignment
- good ways to control access
- models for informed consent and authorship attribution
- methods for anonymization where needed
- inspirational model projects
- a flexible environment for group projects

Transcriptions are important partly because some people prefer to read, but mostly because they make searching and skimming possible. In particular, we badly need

- a well-designed transcription tool that works in web browsers and can read and write distributed audio, video and text
- easy-to-use systems for creating multi-media albums with flexible control over reading, writing, and commenting
- programs for aligning text and audio, and for presenting the results in a way that facilitates multimedia searching, browsing, and skimming.

2.1. Shared goals

There are a large number of other enterprises that share needs with the activities under discussion here. For example, an efficient, flexible, and easy-to-learn web-based transcription system is badly needed for many purposes. And an easy-to-configure system for managing distributed transcription and annotation – assigning tasks,

keeping track of progress, etc. – would be widely used if it were available.

3. How to get there

The Oral History Association has a web site¹ on “Oral History in the Digital Age”, which promises “the latest information on best practices in collecting, curating, and disseminating oral histories.” Content of this type is important as background to the discussion. But the OHA site, with its nearly 100 subsidiary pages on topics like “Metadata: Best Practices for Oral History Access and Preservation,” “Transcripts, Time Coding, and You,” and “File Naming In the Digital Age,” is clearly aimed at a professional audience. The point of this presentation is that with the right environments and the right models, the “collection, curation and dissemination” of such recordings can become a mass-market enterprise.

Many groups and individuals around the world are working on various aspects of this problem. But along with some duplication of effort, there are pieces of the puzzle that no one seems to be working on.

I hope that this workshop will begin a discussion of ways and means, which can continue as an online discussion that informs participants about ideas, techniques and tools, and helps to enlist others in the process.

¹ <http://www.oralhistory.org/ohda-essays/>