# Twitter Crowd Translation — Design and Objectives

**Eduard Šubert**

Czech Technical University
in Prague

**Ondřej Bojar**

Czech Technical University
in Prague

## ABSTRACT

This paper describes our project to support translation of streaming texts on social networks, in particular Twitter. Since machine translation of this type of content is still almost unusable, we rely on volunteers to provide and score the translations. The translations will serve as a testbed and development data for our MT systems tuned for this domain. The project thus serves multiple purposes: From the users' point of view, we would like to provide a smooth access to timely information in foreign languages. From our translators' point of view, we want to provide them with interesting content and material to improve their language skills. Finally, we admit that our project is still primarily a research exercise: As MT researchers, we are interested in learning to handle the specific challenges that this type of content brings. We hope to acquire an interesting collection of data for MT development and to gradually improve our MT processing pipeline for this type of text.

## 1. Introduction

Twitter Crowd Translation (TCT) is our project aimed at the development of an infrastructure for online translation of social media. Through this, we also want to gather relevant training data to support machine translation of such content. We focus on Twitter[1] and the open-source machine translation toolkit Moses. Our project heavily relies on crowdsourcing.

The paper is structured as follows. We first briefly motivate our task and then provide an overview of the system. A section describing some technical details of the implementation follows. Having used the system for a few months in a dry run, we collected some interesting observations on *human* translation of tweets into Section "Translation Aspects of TCT". Finally, we review the related work of machine translation of this content and add our preliminary experience and plans.

[1] http://www.twitter.com/

## 2. Motivation

Social networks have gained tremendous popularity and have successfully replaced many established means of communication. While geographical location of the users has little to no impact on communication, the obstacle of languages used remains.

For stable and long-lasting content, the problem is less severe: services such as Wikipedia have shown that volunteers are able to provide translations into many languages. Machine translation is easy to train on such content and it delivers moderately good results.

In contrast, social networks are used in a streaming fashion, Twitter being the most prominent example. Anybody can contribute a message, which is forwarded to a number of followers. They, in turn, are flooded with messages from sources they select. Given the constant flow of new information, nobody looks back at older messages. Therefore any potential translation needs to be instant.

Providing translation to “streaming networks” is much more challenging. The input is much noisier, significantly reducing MT output quality, and the community is less interested in providing manual translations.

The social motivation of our project is to break the language barrier for streaming social networks. The technological motivation is to advance MT quality by collecting more and better-fit data. What Wikipedia and online MT services manage for stable content, we would like to achieve for streaming networks and casual, unedited text.

Following the open-source culture, we will keep the data and code created in this project fully available to the community.[2]

## 3. Design of TCT

This section provides a high-level overview of the principles we followed when designing TCT.

### 3.1. Overview of Interactions

Our initial intention for TCT was to be as thin layer as possible, to cause minimal disruption. The majority of users would stay within their platform – Twitter in this case.

Therefore, we have designed following system: To select tweets interesting for translation, we manually identified a number of Twitter users and we collect all their tweets. For topicality and attractiveness for our beta-testers, we picked sources from Ukraine that cover the current Russian-Ukrainian conflict in English and Ukrainian.

Once tweets enter TCT, they are immediately sent out to our translators by e-mail. Later the system collects submitted translations and presents them to judges for evaluation. The best translation is tweeted back to Twitter to be published in the most approachable way.

[2] TCT source code is available at http://github.com/cifkao/tct

A few of our beta-testers however demanded the possibility to submit translation through the web. This is in fact in line with the study of Petrović et al. (2010) based on tweets collected in Dec 2009 and Jan 2010: more than 40 % of them were created in the web interface, so many Twitter users don't use any special Twitter application. Having added the option to submit through the web, our original workflow is slightly altered. Figure 1 depicts the current situation.

### 3.2. Benefits for Users

A crucial assumption we make in our project is that there exists a sufficiently large group of users who are happy to contribute for the sake of improving machine translation and therefore the world. This may sound like a ridiculous wish however the online encyclopedia Wikipedia stems from similar ideas and it evolved into a giant information source.

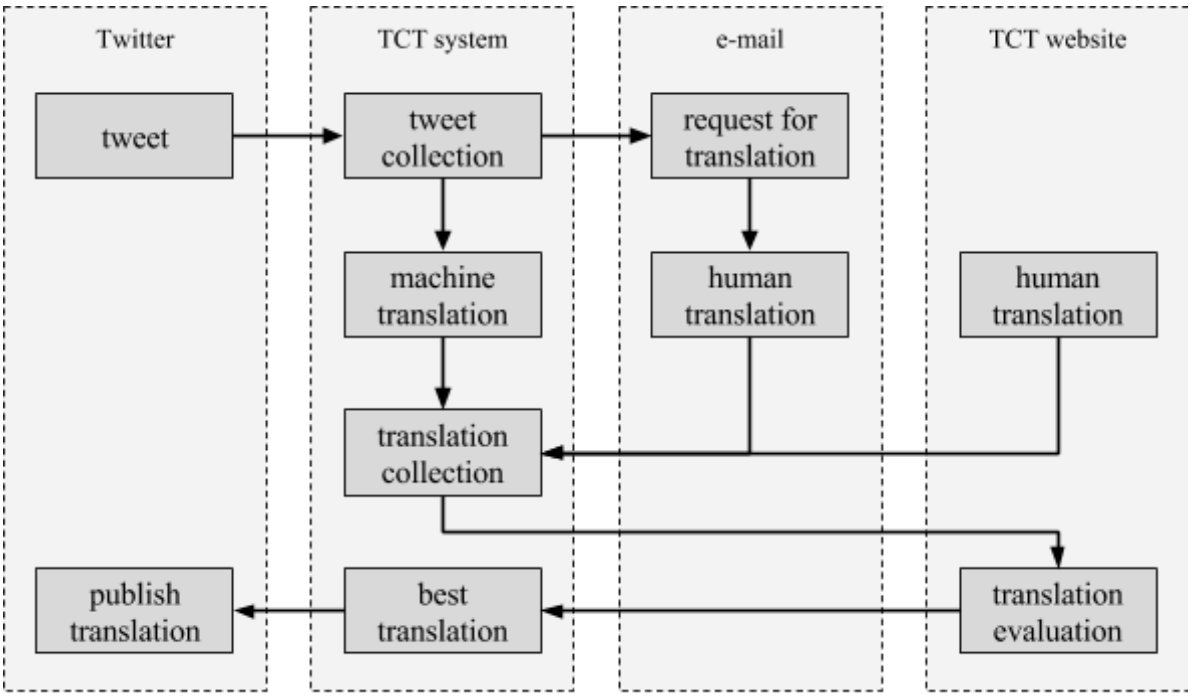


**Figure 1: The TCT workflow in a nutshell**

We speculate that the only incentive for users to contribute to Wikipedia is the prestige. Authors of successful articles may brag about their good work. We will try to deliver similar acknowledgement to our translators for being among the fastest and most accurate.

Therefore, we have implemented a leaderboard for translators. Translators are given an additional space to share information about themselves, including potential contact information, see Figure 2 for an example. This may serve as a registry for prospective customers looking for translation services.

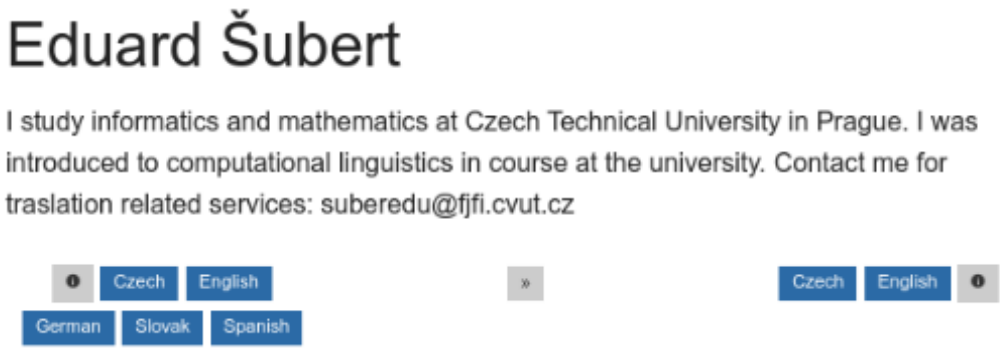


**Figure 2: Sample profile of a TCT translator. The system distinguishes between source and target languages the users is willing to help with**

Aside from gaining reputation as a translator, TCT also offers a field for practice of language skills. We distinguish between languages that translators want to translate *from* and *into*. While still learning a language, people can opt for just reading tweets in it and offer to produce text only in languages they are skill enough at. For very early stages of language learning, users can also contribute by judging translation quality, e.g. from a language they understand only very vaguely.

## 4. Technical Aspects of TCT

This section describes the current TCT workflow in more detail. Technically, the back-end of the application is written in PHP and it is based on the CakePHP framework. The website front-end is built with the aid of Zurb Foundation 5.

In the following, we trace one tweet on its journey through our system from the entry to the publication of its translation.

### 4.1. Collection of Tweets

So far, tweets automatically enter the TCT system only when they are posted by one of the manually selected authors.

In the long term, we definitely plan to add other input ways. For example, we consider searching for tweets based on their originating location to provide translation for a live events. For “on-line events”, following a hashtag would be the most appropriate.

We collect tweets periodically through the Twitter REST API.

### 4.2. Translation of Tweets

Once a tweet enters our system, it is backed up in our database and immediately sent out to all available translators who indicated they can translate the required language pair. We also include our Moses system for some language pairs. Future versions of TCT will need a module for selecting a small subset of translators, mainly for balancing the number of requests sent to each translator based on his or her preferred involvement.

The source language of each tweet is obtained from tweet metadata as Twitter is running its own language detection system. The target language is set manually for each tweet source to reach appropriate audience.

As mentioned above, the initial intention was not to force translators to visit our website in order to submit translations. One reason was to avoid the need for switching into one more application in cell-phone access. Another reason is that we wanted to notify translators that a tweet needs to be translated. Providing the translation upon such a notification should be a one-click activity, to minimize all possible overhead. Therefore, we have based this part of TCT around e-mail as we find it global, simple and since the emergence of smart phones also instant. Translations are simply sent back as replies to our e-mail.

Below is an example of an e-mail sent to our translators. It clearly states the target language and presents the tweet. The URL is a genuine part of the tweet text and it links to an image, video or other supplementary material of the message. The last line labelled "ID" is our hash to identify the source tweet once the translation arrives.

```
Please translate the following post to language: Czech

My congratulations to Ms Dilma Rousseff for her re-election as President
of Brazil. http://t.co/SKamkm5Uci

ID:51d46e56c720c0a20e98e7e628b5806a
```

To fulfill the wishes of our beta testers we have implemented possibility to submit translation through TCT website. As of writing this article, this functionality is still in an early alpha stage.

### 4.3. Translation Evaluation

Since our approach relies on volunteer translators and MT engines, some quality checks need to be part of the processing loop. We follow the common practice and solicit multiple translations. The next step is to select the best translation.

In future, we plan to use automatic quality estimation to either pre-select better translation or to actually perform the selection, see e.g. the system QuEst (Shah et al., 2013) or the shared task on quality estimation (Bojar et al., 2014). The current version of TCT is again limited to manual judging.

Judging via e-mail seemed somewhat cumbersome and we hope to be able to automate this process fully anyway, so we implemented a simple web-based interface.

To ease the work of judges we have first implemented relative pairwise comparison. Presented with the original tweet and a pair of translations, the judge was asked to select the better one or indicate that both are equally good or equally bad (inacceptable).

After some preliminary operation and a long discussion, we moved from relative comparison to absolute judgements. The relative comparison is perhaps easier for the judge but it does not help to decide if the better tweet is *good enough* to be published.

In the current version, the judge is thus presented with the original tweet, one of its translations and the options “hate it”, “dislike it”, “it’s OK”, “like it” and “love it” to indicate his or her attitude towards the translation. We are not entirely satisfied with this solution since we find it more demanding, nevertheless it is the only solution that we could come up with to obtain appropriate data.

With the use of absolute evaluation, it is relatively easy to select the best translation from database and publish it back to Twitter.

### 4.4. Publishing the Best Translation

Another obstacle we had to overcome was publishing the translation back to Twitter. We want to preserve the link between the source tweet and our translation at least informally. Moreover, we want our translations to reach a large audience but we certainly want to avoid being disruptive to users that do not want to see it.

We preserve the link between original tweet and our translation by using @mention, the standard Twitter way to reply to other users. It is technically done by starting the tweet with the other party’s username. This way though, our translations would be visible only to people who follow *both* us and the respective author of the original tweet. This would certainly limit our audience too much.

We therefore use a little technical trick and insert a dot in front of the original author’s name. Figure 2 provides an example where we translated the English tweet by Stefan Füle into Czech. The trick ‘detaches’ our tweet from being a direct reply and makes it visible also to people who follow *only us* and not the original author. Obviously, all our translations are also visible on our profile timeline on Twitter.


**Figure 3: Timeline of our Twitter profile**

## 5. Translation Aspects of TCT

The design of TCT as described above highlights some specifics of (human) translation of tweets.

### 5.1. Understanding Tweets

In contrast to common translation requests (books, news, legal documents, …), tweets are too short to carry sufficient context information themselves. Arguably, the original context of a tweet is not completely unknown. At least the original author, the time and often also the location are recorded. On the other hand, the same author is likely to post tweets about diverse topics from diverse locations, so a more systematic examination of his or her posting history and circumstances would be desirable in order to provide a solid translation.

Our system so far does the exact opposite: we extract individual tweets and send them by e-mail, totally isolated from any available context. The translator is expected to make sense of the tweet without any supportive tools like web search. In a way, we are putting the translator to the position of a machine translation system, with little or no broader understanding of the message. That alone is a very interesting experience for MT researchers.

In the beta run, we picked a topical affair and set our system to follow some manually identified sources from Ukraine. This has already significantly constrained the context so that only very few unrelated items appeared in our collection and also the number of typos, ad-hoc abbreviations and other phenomena typical of social media was considerably smaller.

Some tweets are easy to understand on their own:

- EU-Ukraine Association Agreement provisionally comes into force tomorrow, 1 November. EU welcomes Ukraine! http://t.co/NNkLdh1nHR
- Some tweets use compressed wording that can be seen as vague or imprecise:
- Duma considers creating private military companies for "alternative settlement of armed conflicts outside Russia." http://t.co/RshtPyb06q

Translating this literally, we could cause the impression that the Duma (obviously the Russian parliament, not “male cheetah cub” in Swahili) is indeed planning to establish private companies.

Most tweets however need some rather specific or local world knowledge:

- Terrorists used #Russian-supplied Smerch MRLs against #Ukaine forces in th conflict zone http://t.co/s9ZKP2zsVk |EMPR http://t.co/9ElqFP7qlg
- Col. Lysenko @NSDC_ua spox: truce Memo signed yesterday in Minsk does not work http://t.co/MIYwWrKOi9
- Mykolaiv armored plant handed over 10 APCS to the border guards http://t.co/ANOtAKcUdw via @HromadskeTV http://t.co/mUmfbYKuwf

A likely translator for these tweets in our pool of volunteers will not be an expert in military topics and not a native speaker of English. He or she would thus benefit from additional information looked up and extracted from open sources like Wikipedia, e.g.:

- MRLs = multiple rocket launcher
- truce = ceasefire
- APC = armoured personnel carrier
- “Mykolaiv armored plant” is presumably “346th Mykolayiv Mechanical Armor Repair Plant” in Ukraine, not simply an armoured plant in the city of Mykolaiv.

We are currently working on the specification of such a module for the construction of similar on-demand translation dictionaries.

Sometimes, the picture clarifies the sentence:

- Russian SOF in Ukraine with "Polite People" patch and new gear #CrisisUcrania http://t.co/gqLvuehCof
  - Here, the “Polite People” patch is indeed a cloth badge on the uniform of a mercenary (Soldier of Fortune). (Note the typo in the hashtag.)
- German and soviet officers shaking hands in Brest, both celebrating their joint invasion of Poland. 22 Sept 1939. http://t.co/9EAsFe6cdD
  - Some target languages (e.g. Czech) need to specify if there was one officer per party or more.
- A massive column just outside Donetsk, we re firmly being asked not to film, take pictures.
  - With a picture, we would know if the “column” was of smoke, or if it was a traffic congestion.

And sometimes, the brevity leads to too vague statements:

- @OSCE Chair: so-called elections eastern Ukraine not in line wt Minsk Protocol. Calls for more dialogue, commitments http://t.co/po7W0IFKj1
  - Is it the OSCE chair who calls for more dialogue, or the situation?

Faced with such input sentences and insufficient details about the context, the translator is perhaps going to take a different strategy: to preserve as much ambiguity as possible. Short of any experience in translatology, we are not sure if preserving ambiguity is an established (or even promoted) technique for human translators. As MT researchers, we are not aware of any such maxim in MT whatsoever, but we find this direction of thoughts particularly intriguing and worth exploring.

### 5.2. Guidelines for Producing Translations

As of now, we don’t provide our translators with any requirements, hints or guidelines. During the dry run months, we observed that at least these issues deserve some centralized attention:

- URLs of pictures and detailed sources should be separated from the actual text for translation and reinserted mechanically to the translated tweet – if it fits the length limits.
- Some policy for hashtag translation has to be chosen. Some hashtags are standing away the sentence structure and as such, they can be preserved in the original language while others are part of the syntax and definitely need translation. The policy should say if the hashtag “#Russian” should be included even in the translated sentence where the word Russian has been already translated as ‘ruští’:

    - As #Russian diplomats increase usage of "Novorossiya" important to counter the fake term. The aim is to embed new reality. Old tricks.
    - Jak **ruští** diplomaté stále více používají termín Novorusko (Novorossiya), musíme se proti tomuto falešnému termínu postavit. Cíl je vnutit novou realitu. Staré triky.
- Morphologically rich languages need some policy for hashtags in general. If we decide to translate hashtags to allow for exploration of related tweets in the target language, we may face the problem that hashtags should not be declinated. In this example, the translator decided to use the proper Czech version of the name Lugansk/Luhansk. To preserve the base form of the hashtag, a little trick with the character "|" was used to separate the necessary ending:
    - Mass grave for the dead found during the ceasefire in #Lugansk http://t.co/Bmf8g0bMx6
    - Během příměří byl v **#Luhansk|u** nalezen masový hrob http://t.co/Bmf8g0bMx6
- Usernames like "@someone" may or may not be preceded with an introducing noun (e.g. "uživatel", user) which allows for declination rendering the syntax of the input sentence, the policy should prefer one of the two options.
    - photo by @dondyuk http://t.co/WDj6IfNBDo
    - Fotka od **uživatele** @dondyuk http://t.co/WDj6IfNBDo
- Length limit has to be somehow considered by the translators since the translation is going to be posted as a tweet again. We are planning to cast no technical limit on the translations but to carry out automatic abbreviations if necessary.

## 6. NLP Aspects of TCT

The specifics of social media from the point of view of computational linguistics and natural language processing have been well studied in the past, see e.g. Hachey and Osborne, 2010. Most of this research so far has focused on input normalization and adaptation of NLP tools like taggers, parsers or named entity recognizers for this domain, see e.g. Baldwin et al. (2013) or Bontcheva et al. (2013) for a number of references.

In contrast to this, there seems to be much less research on translation of social media. Microblogs can certainly serve as an interesting source of parallel or comparable corpora, see e.g. Ling et al. (2013), Xing et al. (2013), Rajjem et al. (2013) and Jehl et al. (2012). Gerlach et al. (2013) combine pre-editing and post-editing for user-generated content in a tech forum.

Since the phrase-based approach to MT as implemented e.g. in the Moses toolkit (Koehn et al., 2007) has been shown to successfully circumvent the need for most of linguistic processing, we would like to jump-start MT for tweets in a similar fashion. Nevertheless, we are well aware that most of the mentioned pre-processing tools could bring us an improvement and we plan to gradually add them to future versions of TCT.

Some of such tools are already being developed in open source: Bertoldi et al. 2010 evaluate the utility of confusion networks (code available in Moses) for the recovery from spelling errors and there have been two related MT Marathon 2013 projects: MTSpell (spell checking for machine translation) and SMMTT (Social Media Machine Translation Toolkit, see http://ufal.mff.cuni.cz/mtm13/projects.html). We have already started adapting MTSpell for our needs and our languages of interest.

## 7. Conclusion

We presented TCT, an infrastructure for translation of tweets. As of now, the system is not much more than a playground for researchers in MT. The complete pipeline is nevertheless available for any prospective readers or users: our system follows certain tweet sources, manages registrations of volunteer translators, delivers requests and collects translations from them, operates a manual evaluation of the translations and finally publishes the best translation back to Twitter.

We have already started collecting interesting observations and parallel data necessary to improve MT and MT evaluation for this type of content. We will be gradually automating more and more from our translation pipeline.

## Acknowledgement

The work on this project was supported by the grant FP7-ICT-2011-7-288487 (MosesCore). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

## References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In Proc. of the IJCNLP, pages 356–364, Nagoya, Japan, October 2013. AFNLP.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. Statistical machine translation of texts with misspelled words. In Proc. of HLT/NAACL, pages 412–419, 2010. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In Proc. of WMT, pages 12–58, Baltimore, Maryland, 2014.

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, Proc. of RANLP, pages 83–90. 2013.

Johanna Gerlach, Victoria Porro Rodriguez, Pierrette Bouillon, and Sabine Lehmann. Combining pre-editing and post-editing to improve SMT of user-generated content. In S. O'Brien, M. Simard, and L. Specia, editors, Proc. of MT Summit XIV Workshop on Post-editing Technology and Practice, pages 45–53, 2013.

Ben Hachey and Miles Osborne, editors. WSA'10: Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, Los Angeles, California, 2010. ACL.

Malek Hajjem, Maroua Trabelsi, and Chiraz Latiri. Building comparable corpora from social networks. In BUCC, 7th Workshop on Building and Using Comparable Corpora, LREC, Reykjavik, Iceland, 2013.

Bo Han and Timothy Baldwin. Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter. In Proc. of ACL/HLT Volume 1, pages 368–378, 2011. Association for Computational Linguistics.

Laura Jehl, Felix Hieber, and Stefan Riezler. Twitter translation using translation-based cross-lingual retrieval. In Proc. of WMT, pages 410–421, 2012. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In Proc. of ACL Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June 2007. ACL. http://www.statmt.org/moses/

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. Microblogs as Parallel Corpora. In Proc. of ACL Volume 1, pages 176–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Saša Petrović, Miles Osborne, and Victor Lavrenko. The Edinburgh Twitter Corpus. In Proc. of NAACL/HLT Workshop on Computational Linguistics in a World of Social Media, pages 25–26. 2010. ACL.

Kashif Shah, Eleftherios Avramidis, Ergun Bicici, and Lucia Specia. QuEst – design, implementation and extensions of a framework for machine translation quality estimation. Prague Bull. Math. Linguistics, 100:19–30, 2013.

Haitao Xing, Muyun Yang, Haoliang Qi, Sheng Li, and Tiejun Zhao. Mining Parallel Corpus from Sina Microblog. In Proc. of IALP, pages 99–102, Washington, DC, USA, 2013. IEEE Computer Society.