How to get more data for under-resourced languages and domains?

Andrejs Vasiļjevs

Tilde

Abstract

The explosive growth of digital information on the web enables rapid development of data driven techniques. Significant breakthrough in many areas of language technologies has been achieved. Statistical methods based on huge volume of data have replaced the laborious human work that was required to encode linguistic knowledge. In the new paradigm, the more data you have the better results you get.

However, dependence on data creates new disparities for under-resourced languages and domains. Naturally, smaller language communities produce much less data than speakers of the languages dominating the globe. The same problems occur for language data in narrow domains with their own specific terminological and stylistic requirements.

Essential question is how to get more data for under-resourced languages and domains. Not only innovation needs data but also the collection of data needs innovation. This presentation will briefly discuss some inventive strategies.

When the Web is not enough

Google estimates that 95% of Web pages are in the TOP 20 languages [1]. Although the language identification method used in this estimation is not very reliable, it clearly demonstrates the huge disparity in languages representation on the Web.

Smaller languages often have a complex morphological structure and free word order. To learn this complexity from corpus data by statistical methods, much larger volumes of training data are needed than for languages with simpler linguistic structure.

Shared services from non-sharable data

Motivating users to share their data is a powerful strategy to boost public language resources. In data-driven machine translation most of the public MT systems are built on the parallel texts collected from the web. But a lot of translated data still reside on the local systems of translation agencies, multinational corporations, public and private institutions, and desktops of individual users. TAUS Data Association [2] is an example of successful involvement of the key players in localization industry in sharing their translation memories.

Still in many cases data holders are not able or willing to share their data for competitiveness or confidentiality reasons. New cloud-based services can provide a solution how the community can benefit from restricted data. An example is a machine translation platform being developed through the ICT-PSP project LetsMT! [3]. It provides fully automated self-service for MT generation from user submitted data. As opposed to the traditional sharing platforms users of the LetsMT! system may only upload their data to the online repository. This data is not downloadable and can be used only for the generation and running of statistical models for machine translation. The uploaded proprietary data is not directly exposed or shared. However, the community does benefit from being able to use this data for training and running MT systems. In such a way even small companies and institutions can create their user-tailored MT solutions while contributing to the expansion of online MT training data and a variety of custom MT systems.

Motivating the crowd

The crowdsourcing approach is boosting the acquisition and expansion of language resources. Obviously crowdsourcing needs a crowd. If the total population is in tens or hundreds of millions, a small percentage of active people willing to become involved in crowdsourcing activities make quite a big group. But how about smaller language communities? Boosting enthusiasm or providing monetary based incentives (e.g., Amazon Mechanical Turk) are only temporary solutions to raise the number of participants.

There is a room for innovation to find new motivation schemas. One such example is a successful trial of collaborative translation in Latvia using CTF tool by Microsoft Research [4], and organizing translation competitions in social network.

Use data wisely

Scarcity of data for under-resourced languages and domains is a strong motivation to look for ways to use it more efficiently. For example, a potentially very useful resource could be multilingual comparable corpora – collections of texts about the same or similar topics that are not direct translations. There are several research activities to find efficient methods for collecting and analyzing comparable corpora. FP7 project ACCURAT [5] is researching how to use comparable corpora for statistical MT, but FP7 project TTC - for extraction of multilingual terminology [6].

Profound research is needed to find new ways how to teach computers language tasks. It should be possible to get much better results from much less data than the current data-driven methods. A proof for that is a human child which has fantastic ability to generalize quite limited language information received from the outside world into complete fluency. If we could mimic this ability in online “agents” learning language using web data and interacting with participants from the crowd, this could be a principal solution in closing the technology gap between larger and smaller languages.

References

[1] Daniel Pimienta, Daniel Prado and Álvaro Blanco. 2009. *Twelve years of measuring linguistic diversity in the Internet: balance and perspectives*. UNESCO, Paris.

[2] http://www.tausdata.org

[3] Vasiljevs, Andrejs, Tatiana Gornostay and Raivis Skadins. 2010. *LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation.* Proceedings of the Fourth International Conference Baltic HLT 2010, Riga.

[4] http://blogs.msdn.com/b/translation/archive/2010/03/15/collaborative-translations-announcing-the-next-version-of-microsoft-translator-technology-v2-apis-and-widget.aspx

[5] http://www.accurat-project.eu

[6] http://www.ttc-project.eu