Morphing into the Promised Land

I've been very interested in morphology in translation technology. No, let me put that differently: I've been very frustrated that the translation environment tools we use don't offer morphology. There are some exceptions—such as SmartCat, Star Transit, Across, and OmegaT—that offer some morphology support. But all of them are limited to a small number of languages, and any effort to expand these would require painful and manual coding.

Other tools (e.g., memoQ) have decided that they're better off with fuzzy recognition rather than specific morphological language rules, but that clearly is not the best possible answer either.

So, what's the problem? And what's morphology in translation environment tools about in the first place?

Well, wouldn't it be nice to have all inflected forms of any given word in your source text be associated automatically with the uninflected form that's located in your termbase or glossary, and have that displayed in your terminology search results? And does it feel a little silly to even have to ask that question at a point when it should be a no-brainer to have any given tool provide that service? In case you wondered, the answer to these questions is "Yes, yes, resoundingly yes!"

On the other hand, there's a reason why we're stuck where we are. It happens to be cost. If you really have to manually enter morphology rules for all languages, it quickly becomes a Sisyphean exercise (starting with: "What exactly are *all* languages?"). If you do it just for the "important" languages (which, at least in the eyes of technology vendors, means "profitable"), you end up with the situation we already have with the tools mentioned above.

A few years ago, a group of folks, including myself, had the idea to crowdsource the collection of morphology rules for and with each language-specific group of translators. Once the rules were collected, they could then be integrated into the various technologies. It sounded good, but it was hard to get the project started due to a lack of funds to build the necessary infrastructure and/or the time it would have taken to raise funds, among other issues.

Wouldn't it be nice to have all inflected forms of any given word in your source text be associated automatically with the uninflected form that's located in your termbase or glossary, and have that displayed in your terminology search results?

Enter translation environment tool Lilt with a very cool proposal that may very well be the solution. Lilt's latest version introduces a "neural morphology" engine for all presently supported languages minus Chinese (so: English, Danish, Dutch, French, German, Italian, Norwegian, Polish, Portuguese, Russian, Spanish, and Swedish).

Here is the honest truth, though. When I first read the press release some time ago, I rolled my eyes fondly and thought to myself that the folks from Lilt were just thinking it was wise to throw a little "neural" around while it's hot.

It turns out I was mistaken, however, as I found out when I talked with Lilt's John DeNero, who is the architect of this part of Lilt's system. John tried to explain to me what the system does and why it can make a big difference. It was not so hard to understand the second part, but my feeble untechnical mind had a hard time with the first part.

This column has two goals: to inform the community about technological advances and at the same time encourage the use and appreciation of technology among translation professionals.

ANNOUNCEMENT

NEW ATA PROFESSIONAL LIABILITY INSURANCE BROKER

Alliant is the new broker for ATA-sponsored professional liability insurance. The underwriter is still Lloyds of London. All policies remain in effect and unchanged.

Why choose the ATA-sponsored professional liability insurance?

The policy is specific to the translation and interpreting professions. This is not a one-size-fits-all plan, but a plan that is specific to the work you do.

What does the ATA-sponsored professional liability insurance cover?

- **Broad definition of translation/interpreting services:** Covers activities relevant to a translator or interpreter, including editing, publishing, and proofreading.
- **Coverage for contingent bodily injury and/or property damages:** Covers errors in providing translation/interpreting services that result in bodily injury and/or property damage. These types of claims are typically excluded by generic professional liability policies.
- **Coverage for cyber liability, including HIPPA and HITECH breaches:** Covers breaches related to the provision of professional services in violation or breach of the HIPPA and HITECH Acts.

Visit ata.alliant.com for additional information.

Questions? Contact Alliant at +1-703-547-5777 or ata-questions@alliant.com

(By the way, we always assume that it's us, the less-technically-inclined, who are to be pitied when we don't understand technology. But can you imagine how pitiful life is for the more-technically-inclined who have to speak baby talk when communicating to us?)

An article by Radu Soricut and Franz Och, "Unsupervised Morphology Induction Using Word Embeddings," provides a good summary of the system.[1] It essentially analyzes large monolingual corpora, detects morphological modifications (in theory, they could be any kind of modification; in practice, Lilt focusses on suffixes right now), and classifies them. Since any word is evaluated and also classified *within a context*, the system is able to distinguish between the adverbial ending -ly in English when it encounters "gladly" versus "only." Using the same contextual analysis, the system is also able to make very educated guesses about the morphological transformation of unknown words. (For instance, it might never have encountered "loquacious," but chances are it would assume—correctly—that the adverbial transformation would be "loquaciously.")

This works with every language that uses morphology (therefore excluding Chinese, for instance), provided there is enough corpus material to train the system. The time it takes for a new language to be trained is about two and a half days (on very powerful computers). That's it.

Now, it's not perfect (what is??). John was very open in his assessment about where the system fails. It tends to fail with irregular morphology (it might not recognize "geese" as the plural of "goose" or "well" as the adverbial form of "good"), and there are about 5% of all cases where John felt that the engine should have made a correct judgment but did not.

On the other hand, terminology hits have increased by a third for its users since Lilt introduced the system two weeks ago.

I consider this a quantum leap—in particular because it will not only benefit the large European and Asian languages (where applicable), but the long tail end of other languages as well.

You might say, Lilt covers only a handful of languages, so doesn't that end up being the same thing? The answer to that is (a two-fold) "no." First, you can expect Lilt to continue to add languages, and—even more importantly—the module used to build these neural morphology engines is open-source and available for every translation technology developer online.[2]

Here is what John said about the available engine and its usability:

> Here's our open-source release of the morphology system. It's released as an academic project and doesn't have any formal support, so it's not a product. If someone wanted to use it, they would have to figure it out on their own (though, of course, I'm happy to answer questions).

So, get on it Kilgray, SDL, Atril, Wordfast and, and, and....

It's also very promising that there are other areas where morphological knowledge can be used by a translation system. How about actively changing the inflection of a term that is automatically inserted based on its usage in the source? Or how about changing that inflection when repairing fuzzy matches? Or when repairing machine translation suggestions?

The sky's the limit with this. Be creative!

NOTES

[1] Radu Soricut, Radu, and Franz Och. "Unsupervised Morphology Induction Using Word Embeddings" https://www.aclweb.org/anthology/N/N15/N15-1186.pdf

[2] Oscii Lexicon, https://github.com/oscii-lab/lex.

Jost Zetzsche is the co-author of *Found in Translation: How Language Shapes Our Lives and Transforms the World*, a robust source for replenishing your arsenal of information about how human translation and machine translation each play an important part in the broader world of translation. Contact: jzetzsche@internationalwriters.com.