Other languages

`URL: https://community.almond.stanford.edu/t/other-languages/145`

Question

Hi community!
Wanted to clarify on how to use skills in non-english languages.

As far as I get, to use other languages:

1. one have do deploy “Web+NLP” version of Almond,
2. train custom NLP model which translates text in custom language to the thingtalk code,
3. submit skills with examples in that custom language to the Public ThingPedia,
4. link to NLP server

Questions:
-Did I understand it correctly how to deploy on custom language(above steps)?
-Will the skill in custom language be accepted to Public ThingPedia(which is completely in English now I guess)
-Or any plans to diversify the PublicThingPedia according to language?

嗨社區！
想澄清如何使用非英語語言的技能。

據我所知，使用其他語言：

已經部署了“Web+NLP”版本的 Almond，
訓練自定義 NLP 模型，該模型將自定義語言中的文本轉換為 thingtalk 代碼，
將帶有該自定義語言示例的技能提交給 Public ThingPedia，
鏈接到 NLP 服務器
問題： -
我是否正確理解如何部署自定義語言（以上步驟）？
- 自定義語言的技能是否會被 Public ThingPedia 接受（我猜現在完全是英語） - 或者有任何計劃根據語言使 PublicThingPedia 多樣化？

Answer

Hi shokan,

This is certainly one way to support different languages - but in reality, the limitation is much more core than just training a model.

To support a new language, you need:

- a version of almond-tokenizer 9 that supports that language
- a Genie construct template pack for that language
- translations for all Almond libraries (thingtalk 4, almond-dialog-agent 3, thingengine-core 3)
- translations of Thingpedia metadata (slot-filling questions, canonical forms, confirmation strings)
- translations of Thingpedia primitive templates (dataset.tt files)

- a minimal set of string value sets that include at least tt:location, tt:word, tt:short_free_text and tt:long_free_text

Optionally, you might also want

- special purpose postprocessing / augmentation in Genie
- translations for the Almond frontends
- skill-specific string value sets

Once all the required pieces are in place (with at least one skill) we'll be happy to train a model and deploy it on the public server.

The current status is:

For Chinese (Simplified + Traditional)

- tokenizer is done
- Genie construct templates are available in a branch
- we have a translation of a number of dataset.tt files but we have not uploaded it
- we have not translated the rest of the Thingpedia metadata

For Italian:

- tokenizer is done
- we are working on Genie construct templates (as low priority)
- we don't have any translation of Thingpedia yet (only one skill for testing)

For other language supported by Stanford CoreNLP (Arabic, French, German, Spanish, Russian, Swedish, Danish)

- tokenizer could use CoreNLP, but number and time normalization would need to be provided by a separate package (e.g. HeidelTime 9)

For all other languages, works needs to be done from scratch.

嗨，紹康，

這當然是支持不同語言的一種方式——但實際上，限制不僅僅是訓練模型。

要支持一種新語言，您需要：

杏仁分詞器的一個版本 `9` 支持那種語言
該語言的`Genie`構造模板包
所有 `Almond` 庫的翻譯（`thingtalk 4`,杏仁對話代理 `3`,事物引擎核心 `3`)
`Thingpedia` 元數據的翻譯（空位填充問題、規範形式、確認字符串）
`Thingpedia` 原始模板的翻譯（`dataset.tt` 文件）
一組最少的字符串值集，至少包括 `tt:location`、`tt:word`、`tt:short_free_text` 和 `tt:long_free_text`
或者，您可能還想要

`Genie` 中的特殊用途後處理/增強
`Almond` 前端的翻譯
特定於技能的字符串值集
一旦所有必需的部分都到位（至少具有一項技能），我們將很樂意訓練模型並將其部署在公共服務器上。

目前的狀態是：

中文（簡體+繁體）

標記器完成
Genie 構造模板在分支中可用
我們有一些 dataset.tt 文件的翻譯，但我們還沒有上傳
我們尚未翻譯 Thingpedia 元數據的其餘部分
對於意大利語：

標記器完成
我們正在研究 Genie 構造模板（作為低優先級）
我們還沒有 Thingpedia 的任何翻譯（只有一項技能用於測試）
對於斯坦福 CoreNLP 支持的其他語言（阿拉伯語、法語、德語、西班牙語、俄語、瑞典語、丹麥語）

tokenizer 可以使用 CoreNLP，但數量和時間規範化需要由單獨的包提供（例如HeidelTime 9)
對於所有其他語言，工作需要從頭開始。