How to enroll another language thingpedia device

`URL: https://community.almond.stanford.edu/t/how-to-enroll-another-language-thingpedia-device/380`

Question

Hello. Almond Developers.

I am a graduate student in Korea.
I got to know almond community through the introduction of Mehrad.
My colleague and I are trying to train GenieNLP with KR dataset, and test almond with the model.
We are trying to make kr dataset from scratch, and we have some questions about it.
The questions below are not that easy and simple to reply, but please help us to join this project.

1. How can we make new thingpedia device using Korea(non-English)'s website api?

We are trying to make `manifest.tt`, `dataset.tt` for a new thingpedia device. I saw how to make an English version of the device in Almond Dogs. Can you tell us how we can make tt files from scratch to make a device for Korean?

1. How can we enroll new thingpedia device to local server?

We found that enrolling new device is possible in your dev almond server. But we couldn't find the same skill in local versions of `almond-server`.
(If we make a new model, we want to test the model in almond)

1. How can we make the files needed for making KR dataset?

Now we are referring below page to make kr training dataset.
"https://github.com/stanford-oval/genie-toolkit/blob/master/doc/tutorial-basic.md"
As written in the page, I found out that to create a new device, I needed files like thingpedia.tt, entities.json, etc.
Can you help us to make those files and how can we go through the process?

Thanks!

> 你好。杏仁開發商。
>
> 我在韓國讀研究生。
> 通過介紹 Mehrad，我了解了杏仁社區。
> 我和我的同事正在嘗試用 KR 數據集訓練 GenieNLP，並用模型測試杏仁。
> 我們正在嘗試從頭開始製作 kr 數據集，我們對此有一些疑問。
> 下面的問題不是那麼容易和簡單的回答，但請幫助我們加入這個項目。
>
> 我們如何使用韓國（非英語）的網站api製作新的thingpedia設備？
> 我們正在努力使manifest.tt，dataset.tt新thingpedia設備。我在 Almond Dogs 中看到瞭如何製作英文版的設備。你能告訴我們如何從頭開始製作 tt 文件來製作韓語設備嗎？
>
> 我們如何將新的thingpedia設備註冊到本地服務器？
> 我們發現可以在您的開發杏仁服務器中註冊新設備。但是我們在本地版本的 almond-server.
> （如果我們製作一個新模型，我們想在杏仁中測試模型）

我們如何製作製作KR數據集所需的文件？
現在我們參考下面的頁面來製作 kr 訓練數據集。
“ https://github.com/stanford-oval/genie-toolkit/blob/master/doc/tutorial-basic.md ”
正如頁面中所寫，我發現要創建一個新設備，我需要像thingpedia這樣的文件.tt、entities.json等。
你能幫我們製作這些文件嗎，我們如何完成這個過程？

謝謝！

Answer

> 1. How can we make new thingpedia device using Korea(non-English)’s website api?

The easiest way is to take the English version of a manifest file, and translate all translatable annotations (which are noted by `#_[` instead of `#[`).
The Thingpedia 1 and ThingTalk 1 documentation should also have additional pointers on the syntax for classes (manifest.tt files) and datasets (dataset.tt).

> 2. How can we enroll new thingpedia device to local server?

You should follow the instructions in the Thingpedia testing guide 1. Basically, you make a folder containing a subfolder with your Thingpedia device, with the subfolder named as the Thingpedia device ID, and then you point almond-server to your folder.

Indeed, the easiest is to start from the thingpedia-common-devices 1 repository, which is already set up that way, and also has the Genie Makefiles to generate the dataset and train the model.

> 3. How can we make the files needed for making KR dataset?

Thingpedia.tt is the concatenation of all manifest.tt of all the skills that you want to make a model for. Similarly, to make the dataset.tt you concatenate all the dataset.tt of the individual skills.
You download entities.json from https://thingpedia.stanford.edu/thingpedia/api/v3/entities/all 2

Finally, you need parameter-dataset.tsv. That one you will need to write yourselves because you need the Korean version. The format is a TSV file mapping a `#[string_values]` identifier to a file path with the actual parameters. You can start from the English version which you download using the thingpedia-cli 1. If you use the Genie Makefiles in thingpedia-common-devices or genie-toolkit the English version is prepared automatically.

The big piece though will not be the skill specific files, it will be the domain-independent templates. Those in Genie are at https://github.com/stanford-oval/genie-toolkit/tree/next/languages/thingtalk 1
You basically need to take all the template files under “en”, and translate them to Korean.
We have some work in progress 1 to make the translation a bit less painful (extracting all translatable strings using the gettext 1 workflow). If you want to help us finish that work and make Genie more translatable, that would be wonderful!

For best results, you will also need a language-specific module 1 in Genie. This is do things like split into words, recognize and parse times, dates, numbers in words, convert words to plural and past tense (for languages where that is a thing), and a few more things. If you don’t have the module, you get the default implementation, which splits every ideographic character and recognizes only digits for times/dates/numbers, and doesn’t do any inflection.

1. 我們如何使用韓國（非英語）的網站api製作新的thingpedia設備？
最簡單的方法是獲取清單文件的英文版本，並翻譯所有可翻譯的註釋（用#_[代替 表示#[）。

該Thingpedia 和物語文檔還應該有關於類（manifest.tt 文件）和數據集（dataset.tt）的語法的額外指針。

2. 我們如何將新的thingpedia設備註冊到本地服務器？
您應該按照Thingpedia 測試指南中的說明進行操作 1. 基本上，您使用 Thingpedia 設備創建一個包含子文件夾的文件夾，該子文件夾命名為 Thingpedia 設備 ID，然後將杏仁服務器指向您的文件夾。

確實，最簡單的就是從thingpedia-common-devices開始存儲庫，它已經以這種方式設置，並且還具有用於生成數據集和訓練模型的 Genie Makefile。

3. 我們如何製作製作KR數據集所需的文件？
Thingpedia.tt 是您要為其製作模型的所有技能的所有 manifest.tt 的串聯。類似地，要製作 dataset.tt，您需要連接各個技能的所有 dataset.tt。
您從https://thingpedia.stanford.edu/thingpedia/api/v3/entities/all下載entity.json

最後，您需要參數數據集.tsv。那個你需要自己寫，因為你需要韓文版。格式是一個 TSV 文件，將#[string_values]標識符映射到帶有實際參數的文件路徑。您可以從使用thingpedia-cli下載的英文版本開始. 如果您使用 thingpedia-common-devices 或 genie-toolkit 中的 Genie Makefile，則會自動準備英文版本。

雖然重要的部分不是特定於技能的文件，而是與域無關的模板。Genie 中的那些位於 https://github.com/stanford-oval/genie-toolkit/tree/next/languages/thingtalk
你基本上需要把“en”下的所有模板文件，翻譯成韓文。
我們有一些工作正在進行中 1使翻譯不那麼痛苦（使用gettext提取所有可翻譯的字符串 1工作流程）。如果您想幫助我們完成這項工作並使 Genie 更易於翻譯，那就太好了！

為了獲得最佳結果，您還需要一個特定於語言的模塊在精靈。這是做一些事情，比如拆分成單詞，識別和解析時間，日期，單詞中的數字，將單詞轉換為複數和過去時（對於那些是一件事的語言）等等。如果您沒有該模塊，您將獲得默認實現，它拆分每個表意字符並僅識別時間/日期/數字的數字，並且不進行任何變形。