Sharing User Dictionaries Across Multiple Systems with UTX-S

AAMT Sharing/Standardization Working Group,

Francis Bond,[1] Seiji Okura,[2] Yuji Yamamoto,[3]
Toshiki Murata,[4] Kiyotaka Uchimoto,[1] Michael Kato,[5]
Miwako Shimazu,[6] Tsugiyoshi Suzuki[7]

[1] National Institute of Information and Communications Technology,
[2] Fujitsu Laboratories Ltd., [3] CosmosHouse,
[4] Oki Electric Industry Co., Ltd., [5] Learning Consultant,
[6] Toshiba Solutions Corporation, [7] Cross Language Inc.

AMTA-2008

AAMT Overview

- Introduce a simple dictionary exchange format:
  Universal Terminology eXchange — Simple: UTX-S

  `http://www.aamt.info/english/utx/`

- Use it to swap user dictionaries between systems
  - Improvement on the native system: 44.8% of translations
  - Improvement on a different system: 37.3% of translations

⇒ User dictionaries can successfully be exchanged using a simple interchange format

Motivation

- Using Domain and User dictionaries is the easiest way to increase the quality of machine translation
- So it is important to make user dictionaries
  - Easy to build
  - Easy to share

  Part of the Asia-Pacific Association for Machine Translation (AAMT)'s mission to improve machine translation usability

Previous Work

- Several existing interchange/lexicon formats

  TBX TermBase eXchange
  OLIF Open Lexicon Interchange Format
  UPF Universal PlatForm

- XML-based, powerful formalisms
- Allow most phenomena to be described

X Non-trivial to produce and maintain (even TBX-basic)

AAMT Features of UTX-S

- Extremely lightweight formalism, but extensible
- Based on tab separated values
  - Edit in a spreadsheet
  - Edit in a text editor
- Cannot handle all phenomena
- Covers the most common cases

AAMT Specifications

- A UTX file consists of three parts
  1. A descriptive header (line 1)
  2. A description of the columns (line 2)
  3. The actual entries (tab delimited text)
     entries can be commented out

```
#UTX-S 0.91; en-US/ja-JP; 2008-03-15T10:00:00Z+09:00; copyright: AAMT
#src        tgt        src:pos     src:plural
new         新規の     adjective
fast        高速な     adjective
#prosody    韻律       noun        prosodies
save        保存する   verb
```

AAMT UTX Entry Guidelines

- Add only one translation for each word-pos pair:
  - **the domain-specific best translation**
- Avoid words already in the system dictionary
- Only use the following parts-of-speech:
  {noun|verb|adjective|adverb|properNoun}
- If the pos is unknown then leave it blank: "".
- Detailed guidelines for English and Japanese online:
  `http://www.aamt.info/english/utx/`

AAMT Experiments

1. Test using a domain dictionary converted to UTX-S (+ling)
   lingdic Japanese-English Computational Linguistic Term List
2. Create UTX-S user dictionaries for five systems (+user)
   Translate, and then create user dictionaries based on this
3. Test the user dictionaries on different systems (+other)
   Will a dictionary for system A work with system B?

- Testing done a 147 sentence English document
  "the OLIF Guidelines for Formulating Canonical Forms"

AAMT Systems

Five commercial MT systems were tested:

- LogoVista PRO 2008 Super Pack
- Translation Software ATLAS
- Collaborative Translation Environment: Yakushite.Net
- PC-Transer 2008 Professional (Cross Language Inc.)
- The HON-YAKU 2008 Premium

The results are anonymized as **A**, **B**, **C**, **D** and **E**.

AAMT Sample Translations

- *OLIF Guidelines for Formulating Canonical Forms*

(a) OLIF・正規化形式への定型化の指針 (reference)
"OLIF Guidelines for Formulating Canonical Forms"

(b) 教会法に基づく形式を定式化するためのOLIFガイドライン (MT)
"OLIF Guidelines for formularizing forms based on Canon Law"

(c) 教会法に基づくフォームを定式化するためのOLIFガイドライン (MT+lingdic)
"OLIF Guidelines for formularizing forms based on Canon Law"

(d) 正規化形式を形式化するためのOLIFガイドライン (MT+user)
"OLIF Guidelines for formulating canonical forms"

(e) 基準形を定型化するためのOLIFガイドライン (MT+other)
"OLIF Guidelines for formulating regular forms"

AAMT Domain Dictionary: lingdic

- **lingdic**: open source Ja-En NLP term list
  - 3,527 Japanese head words
  - 4,123 Japanese-English pairs
- mainly used by NLP researchers (and translators)
- We reversed the direction (En-Ja)
  The preferred translation was decided as follows:
  - prefer similar forms
  - prefer common words (web frequency)
  - prefer shorter translations

AAMT User Dictionary

- For each of the five systems
  - translate the text using the domain dictionary
  - add or delete entries to improve the translation
- Added from 17–156 entries (Ave 56)
- Most common term:
  ⟨compound, 複合語, noun⟩ *fukugougo* (in 4 dictionaries)
- A lot of variation (both are OK):
  ⟨string, 文字列, noun⟩ *mojiretsu* "character array"
  ⟨string, ストリング, noun⟩ *sutoringu*

AAMT Shared User Dictionary

- Test if a user dictionary built for one system will also be useful in a different system.
- Swap the dictionaries built above:
  System A uses the dictionary created for system E, B uses the one for A, C uses the one for B and so on.
  - easy to do with UTX-S
- Finally merge all five dictionaries
  - This gives us the upper bound of what can be done with user dictionaries

AAMT Results

System	Dic Size	Percent Change +ling	+user	+other	BLEU Score System	+ling	+user	+other
A	156	-1.4	66.0	38.1	13.8	13.2	21.4	18.2
B	59	-19.7	56.0	20.4	15.4	15.8	18.6	21.1
C	27	-5.4	27.2	36.7	15.5	14.7	17.6	17.0
D	24	-8.2	45.6	32.7	17.2	15.3	20.3	17.8
E	17	2.0	46.9	58.5	12.2	11.7	16.5	16.4
Ave	56.6	-6.5	44.8	37.3	14.8	14.2	18.9	18.1

+ling: results with **lingdic**-EJ
+user: results with a user dictionary built for that system
+other: results with an exchanged user dictionary: A uses E, B uses A, C uses B and so on.

- Merged, corrected dictionary (using system D: 146 entries):
  BLEU = 44.52, an improvement of 27.3 points.

AAMT Discussion – Domain Dic.

- Adding a reversed domain dictionary decreased quality
  6.5% of translation made worse (BLEU -0.6)
- Single word entries degraded existing multi-word entries
  - e.g. *upper case* was 大文字 *oomoji* "capital letters", but changed to 上の格 *ue-no-kaku* "upper (grammatical) case", due to 格 *kaku* "case"
- Reversing the dictionary added errors
  - We need translation frequency, not word frequency

AAMT Discussion – User Dic.

- User dictionaries using UTX-S
  - were simple to build
  - improved translation for 44.8% of sentences (BLEU +4.1)
  - can be compiled in the editor of your choice
- Dictionaries could be shared across systems
- Other systems' user dictionaries (using UTX-S)
  - improved translation for 37.3% of sentences (BLEU +3.3)
- A limited amount of information is still useful

Future Work

- Release the user dictionary conversion tools
- Encourage the production and sharing of dictionaries
  - AAMT validated dictionaries (fee-based)
  - Open dictionaries (unguaranteed)
- Support tuning domain dictionaries for MT
- Cooperation with other projects
  Yakushite.Net, JMDict, Language Grid, . . .

Conclusions

- We have defined a simple user dictionary format
- Used to convert an online glossary to UTX-S
- Produced user dictionaries for five different systems;
  exchanged the dictionaries between systems
- UTX-S can be used to rapidly build dictionaries.
- Customized user dictionaries are effective across systems
  - user dictionaries improved 44.8% of translations
  - shared dictionaries improved 37.3% of translations

AAMT lingdic Sample (UTX-S)

#UTX-S 0.91; en-US/ja-JP; 2008-05-21; copyright: Francis Bond (2008); license: CC-by 3.0

#src	tgt	src:pos
basic lexicon	基本語彙	
co-occurrence dictionary	共起辞書	
collocation dictionary	共起辞書	
concept dictionary	概念辞書	
dictionary	辞書	noun
dictionary form	終止形	noun
generative lexicon	生成的辞書	
idiom dictionary	慣用語句辞書	
idiomatic affix dictionary	連語辞書	
lexicon	辞書	noun
morpho-syntactic dictionary	解析用辞書	
organization of the lexicon	辞書の構成	

AAMT Sample User Dic (UTX-S)

#UTX-S 0.91; en-US/ja-JP; 2008-05-30;

#src	tgt	src:pos
canonical	形式化	adjective
canonical form	正規化形式	noun
compound	複合語	noun
compound noun	複合名詞	noun
convention	規定	noun
enter	入力する	verb
formulate	形式化する	verb
multiple-word	複数単語からなる	adjective
SAP	SAP	noun
spelling convention	一般的なスペル表記	noun
string	文字列	noun
the head	先頭の	adjective
usu.	通常	adverb