Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks

Yoichiro Hasebe
Introduction

TED Corpus Search Engine (TCSE)
http://yohasebe.com/tcse

- Capable of searching TED Talks and show text/audio/video segments that match the input string
- Not only English transcripts but also their translations in 15 languages are available
- Designed and implemented on the usage-based model of language
About TED

TED (http://ted.com)

- Technology, Entertainment, and Design
- TED and TEDx conferences are held worldwide to share ideas worth spreading
- Speakers include artists, researchers, politicians, etc.
- Talk data are distributed under Creative Commons License (by-nc-nd 3.0)
Transcripts and translations

• More than 18,000 transcripts of TED talks are available online
• There are many volunteers transcribing, translating, and reviewing the talk text

Translation stats
• 107 languages
• 19,807 translators
• 70,021 translations

(http://www.ted.com/participate/translate)
Limitations of official TED web system

- Text search functionality is not very sophisticated (no POS and lemma searches)
- Comparison between transcript and translation is not possible
- Comparison between translations is not possible
- Video/audio starts from the beginning (not from the segment that matches the input string)
TED as corpus

Cons

• It is not balanced.
• Token size is about 4,500,000 (not comparable to corpora such as BNC and COCA).
• It contains English by many non-native speakers.

Pros

• It is provides rich audio/video data.
• It can be considered as a corpus of presentation text.
• It contains wide varieties of English in the world.
Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks

http://yohasebe.com/tcse

About TCSE
TCSE is a search engine specializing in exploring transcripts of TED Talks. It has been created for educational and scientific purposes. TCSE uses data provided by TED under the Creative Commons BY-NC-ND license, but it is not an official service of TED.

Using and Citing TCSE
TCSE is created by Yoichiro Hasebe and made available free for non-commercial educational and scientific use. Please cite one of the following when you publish work which utilizes TCSE. (The content of the manual is outdated. It will be updated soon!)


Change Log
- 2015/02/21
  Tentative release of 2.0 with many bug fixes. Translation text (partially) in 15 languages (= left-to-right languages to which more than 1,500 talks have been translated so far) are available for search now! (The advanced search in Japanese had to be removed as trade-off though.)
- 2014/11/28

About Advanced Search
Advanced search is available only in English.
- List of English POS tags
  POS keys are specified either fully (e.g.,  ) or partially (e.g.,  ).
  An advanced search query string cannot consist only of POS keys.

Advanced Search Syntax

- Lemma
  \[\text{[LEMMA]}\]

- Part of Speech
  \[\text{[POS]}\]

- Surface + Part of Speech
  \[\text{SURFACE(POS)}\]
  (with no spaces in-between)

- Lemma + Part of Speech
  \[\text{[LEMMA] [POS]}\]
  (with no spaces in-between)

- Logical Disjunction (OR)
  \[\text{A | B}\]

- Wild Card
  \[\text{\_}\]
Specifications of TCSE

Text units of different sizes

- **segments**
- expanded segments
- paragraphs

<table>
<thead>
<tr>
<th>#</th>
<th>Talk ID</th>
<th>Line [Position]</th>
<th>Time [Total]</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2194</td>
<td>149 [0.6]</td>
<td>09:28 [15:18]</td>
<td>when we know not only which pathways are disrupted, but how,</td>
</tr>
<tr>
<td>2</td>
<td>2193</td>
<td>90 [0.31]</td>
<td>05:23 [16:56]</td>
<td>I didn't have an answer then, but I do today,</td>
</tr>
<tr>
<td>3</td>
<td>2193</td>
<td>206 [0.72]</td>
<td>12:20 [16:56]</td>
<td>Maybe not as harshly, but we all do it,</td>
</tr>
<tr>
<td>4</td>
<td>2178</td>
<td>108 [0.33]</td>
<td>04:34 [14:39]</td>
<td>depends not on you, but on what you're trying to learn.</td>
</tr>
<tr>
<td>5</td>
<td>2178</td>
<td>209 [0.64]</td>
<td>09:07 [14:39]</td>
<td>is that listening to Mozart can make you not only cleverer but healthier, too.</td>
</tr>
<tr>
<td>6</td>
<td>2177</td>
<td>130 [0.41]</td>
<td>07:58 [18:50]</td>
<td>not receiving the support they needed, but increasingly isolated.</td>
</tr>
<tr>
<td>7</td>
<td>2176</td>
<td>8 [0.02]</td>
<td>00:20 [15:32]</td>
<td>This is not inevitable, but overcoming it requires diving deep</td>
</tr>
<tr>
<td>8</td>
<td>2176</td>
<td>175 [0.64]</td>
<td>10:15 [15:52]</td>
<td>not just the march, but the capacity signaled by that march, seriously.</td>
</tr>
<tr>
<td>9</td>
<td>2176</td>
<td>214 [0.78]</td>
<td>12:31 [15:52]</td>
<td>It's not the same game, but the differences are instructive.</td>
</tr>
<tr>
<td>10</td>
<td>2174</td>
<td>122 [0.8]</td>
<td>08:03 [15:34]</td>
<td>And that happens not only in Congo but also in many other conflict zones.</td>
</tr>
</tbody>
</table>
Specifications of TCSE

Text units of different sizes

• segments
• expanded segments
• paragraphs

“not A but B” construction

[not] * but

975 segments

4189 expanded segments
Specifications of TCSE

Text units of different sizes

- segments
- expanded segments
- paragraphs
Specifications of TCSE

Regular search in English transcripts
1 (a) might as well
   (b) as if
   (c) far from

Regular search in translations (Japanese examples)
2 (a) かもしれない
   (b) まるで
   (c) 程遠い
Specifications of TCSE

Lemma search
3 (a) [make] sense
   (b) [know] better
   (c) [happen] to

Lemma and POS search
4 (a) [remember] to
   (b) [remember] {v}
5 (a) [help]{n}
   (b) [help]{v}  (note: no space between [] and {})

Doshisha University
Specifications of TCSE

Wild card and logical disjunction
6  [not] * but
7  [also] long as

Specification of segment onset position
8  (a) ^ again
   (b) ^ still
   (c) ^ now
# Statistics of talks in TCSE

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of talks</td>
<td>1,828</td>
</tr>
<tr>
<td>Total playing time of talks</td>
<td>396 hours</td>
</tr>
<tr>
<td>Total number of segments</td>
<td>511,923</td>
</tr>
<tr>
<td>Total number of expanded segments</td>
<td>220,565</td>
</tr>
<tr>
<td>Total number of word tokens</td>
<td>4,567,505</td>
</tr>
<tr>
<td>Total number of word types</td>
<td>80,790</td>
</tr>
<tr>
<td>Mean length of talks</td>
<td>13 minutes</td>
</tr>
<tr>
<td>Mean number of words</td>
<td>2,499</td>
</tr>
<tr>
<td>Mean words per minute</td>
<td>192</td>
</tr>
</tbody>
</table>

(as of February 21, 2015)
Transliterations and Implementations of an Online Corpus of Presentation Transcripts of TED Talks

Translations

マチ・リカール: 愛他性に導かれる生き方
愛他性とは何でしょうか？単に言えば、自分以外の人達の幸せを願うことです。そして、幸福学の研究者で仏教僧侶であるマチ・リカールが言うには、愛他性は短期的なものであれ長期的なものであれ、また仕事に関することであれ暮らしに関することであれ、何か決断を下す時に重要な指針として働くものです。

68 04:33 and the long term of the environment. そして環境という長期的なものです
67 04:37 When the environmentalists speak with economists. 環境活動家が経済学者と話をする
68 04:39 It's like a schizophrenic dialogue, completely incoherent. 統合失調推進の会話になってしまうが全く通じません
69 04:42 They don't speak the same language. 彼らは話す言語が違っているんです
70 04:45 Now, for the last 10 years, I went around the world. この10年私は世界中を回って旅の
71 04:49 meeting economists, scientists, neuroscientists, environmentalists, 経済学者、科学者、神経学者、環境活動家
72 04:53 philosophers, thinkers in the Himalayas, all over the place. 哲学者、ヒマラヤの住民たち、全ての場所
73 04:57 It seems to me, there's only one concept. 先程の3つの時間尺度の間を取り持てるのは
74 05:01 that can reconcile those three time scales. たった1つの概念しかないように思えます
75 05:04 It is simply having more consideration for others. それ故に、他者をもっと大事にするということです

ジョン・マクホーﬀ 「テキスト・メッセージが言語を殺す （なんてね！）」

<table>
<thead>
<tr>
<th>Surface</th>
<th>and</th>
<th>look</th>
<th>at</th>
<th>what</th>
<th>language</th>
<th>really</th>
<th>is</th>
<th>。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemma</td>
<td>and</td>
<td>look</td>
<td>at</td>
<td>what</td>
<td>language</td>
<td>really</td>
<td>be</td>
<td>—comma—</td>
</tr>
<tr>
<td>POS</td>
<td>cc</td>
<td>vb</td>
<td>in</td>
<td>wp</td>
<td>nn</td>
<td>rb</td>
<td>vb</td>
<td>。</td>
</tr>
<tr>
<td>Freq</td>
<td>132,819</td>
<td>4,367</td>
<td>16,236</td>
<td>23,713</td>
<td>915</td>
<td>8,763</td>
<td>52,336</td>
<td>247,679</td>
</tr>
<tr>
<td>PerMill</td>
<td>29,079</td>
<td>969</td>
<td>3,555</td>
<td>5,192</td>
<td>200</td>
<td>1,919</td>
<td>11,458</td>
<td>54,226</td>
</tr>
</tbody>
</table>

Transcript and Translation

English | and look at what language really is, ▲
Chinese, Simplified | 好好想想語言究竟是怎麼一回事。 ▲
Chinese, Traditional | 看着語言到底是怎麼一回事 ▲
Dutch | om te kijken wat taal eigenlijk is. ▲
French | et examiner ce qui est vraiment la langue. ▲
German | und untersuchen, was Sprache wirklich ist. ▲
Italian | e vedere che tipo di linguaggio è veramente, ▲
Japanese | 言語の本質を見る必要があります ▲
Korean | 언어의 본질을 찾는 것도 필요하다. ▲
Portuguese, Brazilian | e analisarmos o que a língua realmente é, ▲
Romanian | ști să înțelegem ce este fapt limbajului. ▲
Russian | нужно понять, что такое есть сам язык, ▲
Spanish | y observar qué es en realidad el idioma. ▲
# List of translation languages

15 languages
(languages written L-to-R to which more than 1,500 talks have been translated)

<table>
<thead>
<tr>
<th>Language</th>
<th>Talks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>1,567</td>
</tr>
<tr>
<td>Chinese, Simplified</td>
<td>1,697</td>
</tr>
<tr>
<td>Chinese, Traditional</td>
<td>1,634</td>
</tr>
<tr>
<td>Dutch</td>
<td>1,448</td>
</tr>
<tr>
<td>French</td>
<td>1,572</td>
</tr>
<tr>
<td>German</td>
<td>1,527</td>
</tr>
<tr>
<td>Italian</td>
<td>1,674</td>
</tr>
<tr>
<td>Japanese</td>
<td>1,600</td>
</tr>
<tr>
<td>Korean</td>
<td>1,697</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1,026</td>
</tr>
<tr>
<td>Portuguese, Brazilian</td>
<td>1,593</td>
</tr>
<tr>
<td>Romanian</td>
<td>1,630</td>
</tr>
<tr>
<td>Russian</td>
<td>1,744</td>
</tr>
<tr>
<td>Spanish</td>
<td>1,667</td>
</tr>
<tr>
<td>Turkish</td>
<td>1,495</td>
</tr>
</tbody>
</table>

**Note:** some translations released by TED are not imported to TCSE for technical reasons.
Usage-based model of language

Basic theoretical concept behind TCSE


The usage-based thesis holds that the mental grammar of the language user ... is formed by the abstraction of symbolic units from situated instances of language use: an utterance. (Evans 2007: 216-217)
Schemas and instances/exemplars
Getting (good) exemplars from corpus

For linguistic research
TCSE provides “situated” instances of expressions
→ especially important in:
  • cognitive linguistics
  • discourse analysis

For language learning/teaching
TCSE offers readily available real samples of:
  • usages of words, phrases, constructions, etc.
  • possible translations of particular expressions
Conclusion

TED Corpus Search Engine (TCSE)

• available online at http://yohasebe.com/tcse
• searches more than 1,800 TED talk transcripts in English and translations in 15 languages
• designed and implemented based on usage-based model of language

Thank you!
Yoichiro Hasebe (Doshisha University, Kyoto, Japan)
yohasebe@gmail.com