

Engineering Part IIB: Module 4F11

Speech and Language Processing

Lecture 11: Introduction to Statistical Machine Translation

Bill Byrne

Lent 2014



Cambridge University Engineering Department

Machine Translation

Machine Translation is the full or partial automation of translation as carried out by humans. This is a difficult objective to define in absolute terms, since human translation often involves multiple overlapping tasks, for example

- ▶ Translation
- ▶ Interpretation
- ▶ Summarization
- ▶ Transliteration
- ▶ Glossing
 - ▶ Syntactic and Semantic Analysis
 - ▶ Commentary

All of these tasks require human intelligence, sensitivity, and understanding clearly beyond the state-of-the-art in NLP and artificial intelligence.

However translated texts are readily available in a variety of domains

- ▶ Religious texts
- ▶ Governmental proceedings:
EU, UN, multilingual countries ...
- ▶ Business documents:
localization of product manuals, ...
- ▶ Literary texts
- ▶ News sources:
multilingual and multinational
- ▶ Consumer: tourism, shopping, ...
- ▶ Military and humanitarian efforts

For example ...

The Eadwine Psalter, Canterbury, Mid-Twelfth Century



...Three Latin versions of the Psalms laid in parallel columns are integrated with a Latin commentary, Old English and Anglo-Norman translations written between the lines and in the margins. Each Psalm opens with a magnificent drawing inspired by its text the waters of Babylon, illustrates Psalm 136 shown here...

The Cambridge Illuminations,
The Fitzwilliam Museum ¹

¹<http://www.fitzmuseum.cam.ac.uk/gallery/CambridgeIlluminations>

European Union Parliamentary Proceedings

Rules of Procedure of the European Parliament – Rule 138

01/02/2006 09:55 PM



▶ TITLE VI : SESSIONS

CHAPTER 3 : GENERAL RULES FOR THE CONDUCT OF SITTINGS

Rule 138 : Languages

1. All documents of Parliament shall be drawn up in the official languages.
2. All Members shall have the right to speak in Parliament in the official language of their choice. Speeches delivered in one of the official languages shall be simultaneously interpreted into the other official languages and into any other language the Bureau may consider necessary.
3. Interpretation shall be provided in committee and delegation meetings from and into the official languages used and requested by the members and substitutes of that committee or delegation.
4. At committee and delegation meetings away from the usual places of work interpretation shall be provided from and into the languages of those members who have confirmed that they will attend the meeting. These arrangements may exceptionally be made more flexible where the members of the committee or delegation so agree. In the event of disagreement, the Bureau shall decide.

Where it has been established after the result of a vote has been announced that there are discrepancies between different language versions, the President shall decide whether the result announced is valid pursuant to Rule 164(5). If he declares the result valid, he shall decide which version is to be regarded as having been adopted. However, the original version cannot be taken as the official text as a general rule, since a situation may arise in which all the other languages differ from the original text.

Last updated: 17 October 2005

Legal notice



European Union Parliamentary Proceedings

German

Offensichtlich bedeutet die Erklärung von Herrn Fischler vom Wochenende eine Änderung der Haltung der Kommission.

Ich begrüße diese Änderung , denn er sagte , daßer britishes Rindfleisch essen würde und daßdas Einfuhrverbot insbesondere aus wirtschaftlichen und politischen Gründen verhängt wurde.

English

It would appear that a speech made at the weekend by Mr Fischler indicates a change of his position.

I welcome this change because he has said that he will eat British beef and that the ban was imposed specifically for economic and political reasons.

French

Il semblerait en effet que M. Fischler ait changé de position dans un discours prononcé au cours de ce week-end.

Je me félicite de ce changement, car il a dit qu'il mangerait du boeuf britannique et que l'interdiction avait été décrétée spécifiquement pour des raisons économiques et politiques.



The DGT Multilingual Translation Memory of the Acquis Communautaire: DGT-TM

As of November 2007, the European Commission's Directorate General for Translation (DGT) made publicly accessible its multilingual Translation Memory for the Acquis Communautaire (the body of EU law) - a collection of parallel texts (texts and their translation, also referred to as bi-texts) in 22 languages. (<http://langtech.jrc.it/DGT-TM.html>)

4) Statistics for the DGT Translation Memory

The *DGT Translation Memory* is currently available in 22 languages. The following table shows the coverage, expressed in the total number of translation units available for each language:

Language	Language code	Number of units
English	EN	2 187 504
Bulgarian	BG	708 658
Czech	CS	890 025
Danish	DA	433 871
German	DE	532 668
Greek	EL	371 039
Spanish	ES	509 054
Estonian	ET	1 047 503
Finnish	FI	514 868
French	FR	1 106 442
Hungarian	HU	1 159 975
Italian	IT	542 873
Lithuanian	LT	1 126 255
Latvian	LV	1 120 835
Maltese	MT	1 021 855
Dutch	NL	502 557
Polish	PL	1 052 136
Portuguese	PT	945 203
Romanian	RO	650 735
Slovakian	SK	1 065 399
Slovene	SL	1 026 668
Swedish	SV	555 362

Size of DGT's Translation Memory expressed as the total number of translation units per language for each of the 22 official EU languages.

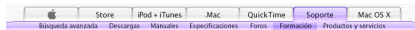
The number of *aligned translation units* differs for each language pair.



Product Localization

Apple - Formación

01/02/2006 10:23 PM



El servicio de formación de Apple

El servicio de formación de Apple se enorgullece de ofrecer cursos de formación de alta calidad en todo el mundo sobre productos y tecnologías de Apple. Estos cursos presenciales combinan conceptos y clases teóricas con laboratorios y ejercicios prácticos. Los cursos de autoestudio ofrecen sesiones dinámicas e interactivas en las que aprenderás diferentes temas.

<p>Mac OS X y Mac OS X Server</p> <p>Apple pone a disposición de todos (desde particulares a administradores de sistemas) completas oportunidades de formación sobre Mac OS X y Mac OS X Server.</p> <p>Más detalles</p>	<p>Formación sobre aplicaciones profesionales</p> <p>Aprende todo lo necesario de tecnologías multimedia de Apple (Final Cut Pro HD, DVD Studio Pro, Shake y Logic) en nuestros cursos presenciales o de autoestudio.</p> <p>Más detalles</p>
<p>Formación para técnicos AppleCare</p> <p>Prepárate para los exámenes de certificación de Apple y aprende a resolver problemas de software y hardware de Apple.</p> <p>Más detalles</p>	<p>Instructor profesional</p> <p>Los cursos de desarrollo y formación profesional te ayudan a aprovechar al máximo las tecnologías en el aula y a motivar a tus alumnos hacia el éxito.</p> <p>Más detalles</p>

[Inicio](#) > Formación

Visita el Apple Store en [España](#)

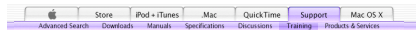
800 130 1000

[Mapa del sitio](#) | [Contacta con nosotros](#) | [Condiciones de uso](#) | [Política de privacidad](#)

Copyright © 2005 Apple Computer, Inc. Todos los derechos reservados.

Apple (UK and Ireland) - Training

01/02/2006 10:22 PM



Apple Training

Apple Training is proud to offer high-quality training on Apple products and technologies worldwide. The leader-led offerings balance concepts and lectures with hands-on labs and exercises. The self-paced courses offer interactive, dynamic sessions that walk you through various topics.

<p>Mac OS X and Mac OS X Server</p> <p>Apple offers everyone – from individuals to system administrators – comprehensive training opportunities on Mac OS X and Mac OS X Server.</p> <p>Learn More</p>	<p>Pro Apps Training</p> <p>Learn about Apple's digital media technologies (Final Cut Pro HD, DVD Studio Pro, Shake and Logic) in instructor-led or self-paced courses.</p> <p>Learn More</p>
<p>AppleCare Technician Training</p> <p>Prepare for Apple Service Certification exams while learning to troubleshoot and resolve software issues and repair Apple hardware products.</p> <p>Learn More</p>	<p>Professional Educator</p> <p>Apple professional development and training helps you make the best use of classroom technologies and raise student achievement.</p> <p>Learn More</p>

[Home](#) > Training

Quick Clicks

Programmes

- [Apple Certification](#)
- [Mac OS X Learning Centres](#)
- [Apple Developer Connection](#)
- [Apple Connect](#)

Training

- [Apple Authorized Training Centre Programme](#)
- [Apple Authorized Training Centre Locations](#)
- [Training Resources](#)

Training Resources

- [DVDs and CDs](#)
- [Books](#)

Software Training

- [Mac OS X \(New Users\)](#)
- [Mac OS X \(IT Professionals\)](#)
- [Mac OS X Server](#)
- [DVD Studio Pro](#)
- [Final Cut Pro HD](#)
- [Logic Pro](#)
- [WebObjects](#)

Visit the Apple Store [online](#) or at [retail](#) locations.
0800 039 1030 (UK) / 1800 92 38 98 (Republic of Ireland)

[Site Map](#) | [Contact Us](#) | [Terms of Use](#) | [Privacy Policy](#)
Copyright © 2005 Apple Computer, Inc. All rights reserved.



News Sources with Multilingual and Multinational Audiences

5a-5e-cs | RSS
26/02/2008 17:19

BBC Home News Sport Radio TV Weather Languages **ВОСН**

BBC RUSSIAN.com
Без графика | О сайте | Обратная связь.

Главная
В мире
Россия
Экономика
Наука и техника
Люди
Культура
Британия
Аналитика
Ваша история
Мир в карте
Learn English
Радио
Видео
Партнеры

В мире
Обновлено: вторник, 26 февраля 2008 г., 17:05 GMT 20:05 MSK

Видео
Китайец задержан на иммиграции в Бордо
Видеополета по США: чужие страны
Американец выиграл \$275 млн.
Фидель Кастро призвал кубинского лидера в риторических

ДРУГИЕ ВИДЕОСЮЖЕТЫ >>

Актуально
Что такое RSS?
5a-5e-cs на ДРУГИХ ЯЗЫКАХ
УКРАИНСЬКА
АЗЕРБАЙДЖАН
РУССКА
ДАЛЕКО НЕ
ОСТАЛЫЙ >

ДРУГИЕ НОВОСТИ
Українська преса: Парламентальні повороти в череві
Пост економіки США "розкривається"
"Сандбу" обрався в Індії: посилали тисячі

ІЗБРАННЕ
Новий сервіс BBC Russian Select: що це таке і зачим його нуруно?
Наше відео Саміх сюжетів світової BBCRussian.com

Українська преса: Коалиція зможе прористися з жовтій
Більш торбан сенатора Обаму
Новий свідчення о ядерній програмі Ірану

Новий сервіс BBC Russian Select: що це таке і зачим його нуруно?
Наше відео Саміх сюжетів світової BBCRussian.com

Русская служба Би-би-си - Информационные услуги
 Новости по почте Для мобильных устройств Новостной канал (RSS)

О сайте | Обратная связь | Партнеры | Погода | Конфиденциальность

Listen live
 Listen live
 Listen live



26/02/2008 17:19

BBC NEWS Home News Sport Radio TV Weather Languages **Search**

Low graphics | Accessibility help

Europe Today
 LISTEN LIVE

NEWS Front page
 Africa Americas Asia-Pacific Europe Middle East South Asia UK
 England Northern Ireland Scotland Wales UK Politics Education Magazine Business Health Science/Nature Technology Entertainment

UK
 Last Updated: Tuesday, 26 February 2008, 17:16 GMT

Top terror recruiter found guilty
 A man said to be one of the most influential recruiters for Islamist extremism in the UK is found guilty at a major trial.
 The radicaliser and the bombers
 Hate speech: What they said
 UK 'terror camps' revealed

Bus stop killer gets whole life term
 A man who killed two students and tried to kill a third must spend the rest of his life in prison, a judge orders.

Lib Dem ordered out of EU debate
 Lib Dem frontbencher Ed Davey is ordered out of the Commons after protest over EU treaty debate.

Nick Robinson
 Oh dear, it seems I've angered a growing number of MPs

Anti-depressants 'do not work'
 Jersey 'abuse cover-up' claim
 Mars vehicle model tested

More Video / Audio News
 SHOW SHOW

VIDEO AND AUDIO NEWS
 TV AND AUDIO PROGRAMMES
 MOST POPULAR VIDEO AND AUDIO

MORE FROM THE UK
 Iraq minutes 'should be released'
 UK work applicants near 800,000
 Hospital staff 'sued in porn probe'
 Average water bill rising by 5.8%
 Truancy rate 'highest since 1997'
 Egg complaints come to the boil
 100,000 miss first-choice school
 Materialism 'damages' children
 Cash plea for die-at-home plan
 Two quizzes over fatal stabbing

MAGAZINE
 For the love of bricks
 How can a dog sniff through concrete?
 My baby just loves to dance

UK NEWSPAPERS
 Papers say 'the drugs don't work'

OTHER TOP STORIES
 Super-casino proposal is ditched
 Jersey 'abuse cover-up' is denied
 Anti-depressants' 'little effect'
 MPH 'expanses review speeded up'
 NHS 'did not keep files on Osama'
 Missing girl police 'fear worst'
 Bank warns of economic risks

ALSO IN THE NEWS
 Small tawny owl turns heads after being rescued from pipe

SPORT HEADLINES
 Hicks denies sale claim
 Football: Moyes' European focus
 Athletics: Williamson makes team

HOUSTON
 Six-year mystery
 What happened to bright, vivacious Milly Dowler?

HAVE YOUR SAY
 66 Companies know that the gateway to your money is through your children
 Emma, Katterling

MOST POPULAR STORIES
 MONDAY
 Pilot sacked after fly-by stunt
 TUESDAY
 Active in first biofuel film
 SATURDAY
 Mandatory DNA database rejected
 FRIDAY
 Claim UK troops 'executed' Iraqi
 THURSDAY
 New airport flyt paths proposed
 UK WEATHER
 Select a city: [input] GO

RELATED BBC SITES
 SPORT WEATHER
 ON THIS DAY EDITORS' BLOG

Вернуться ^^^

These Sources Suggest Both a Demand and an Opportunity for MT

Government and commercial bodies are producing ever increasing amounts of translated speech and text.

- ▶ Production is driven by political and economic necessity
- ▶ Dissemination is aided by the internet and other broadcast media

This is good for research and development in MT technology

- ▶ Useful products are needed
- ▶ Technology can be put to use – and evaluated – in real applications
- ▶ Data is being produced for use in system development

There is clearly a market for translation services

- ▶ Demand is currently met by (relatively) well-paid professional translators
- ▶ High-quality translation are very expensive and can be difficult to obtain
- ▶ Even low-quality translation services are expensive
 - ▶ Partial automation - *Computer Aided Translation* - may have economic value
- ▶ Commercial translation services are ubiquitous on the web


Why is Machine Translation Difficult ² ?

Variations within language pairs and translations domains challenge even bilingual humans

- ▶ Domain and Genre
- ▶ Word sense
- ▶ Morphology
- ▶ Word Order
- ▶ Idiomatic expressions
- ▶ Language Specific Issues:
 - Chinese - word segmentation
 - Arabic - tokenization
- ▶ Translation vs. Transliteration
 - named entity identification

Constructing models for use in statistical machine translation is computationally difficult

- ▶ Translation does not proceed in left-to-right fashion (as in automatic speech recognition) due to word reordering between language pairs
- ▶ Large amounts of monolingual and bilingual text are needed for parameter estimation

²Much of the material in this section was taken from Jurafsky and Martin, Chapter 21 



Translation Should Respect Document Domain and Genre

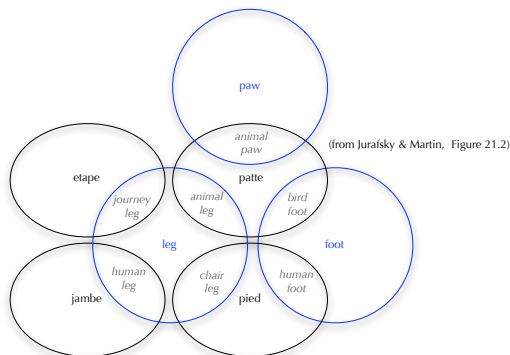
Religious texts, legal documents, business correspondence, technical documents, news sources, etc. are sometimes referred to as *sublanguages*

- ▶ Sublanguage domains differ across any of a range of dimensions
 - ▶ Specialized vocabularies
 - ▶ Stylistic differences, e.g. active vs. passive voice
 - ▶ Use of headlines, passage and chapter numbers, web addresses, proper names, ...
 - ▶ Variable sentence lengths, e.g. weather forecasts vs. news stories
- ▶ The same translation system cannot be used for all genres
- ▶ Text from one genre may not be suitable for building systems in another genre

Word Sense and Context: Translation Must Depend on Word Sense

Ambiguous words should be translated based on their *sense* : e.g. bank, plant, leg, ...

- ▶ Should an instance of *leg* be translated as *etape* , *pied* , *patte* , or *jambe* ?



Fortunately word sense is often (easily) determined by *context*, even for idioms

... *you're just pulling my leg* ... (English) \Leftrightarrow ... *mi stai prendeno per il naso* ... (Italian)

Morphology

Morphological variation across languages leads to modeling difficulties in translation

- ▶ Morphological analysis must be applied prior to processing
- ▶ Surface variability leads to sparsity of individual tokens in text translations

Example of Arabic tokenization and morphological analysis

. وينتج المفاعل البلوتونيوم اللازم لتصنيع القنبلة الذرية .

[and-produces] [the-reactor] [the-plutonium] [the-required] [to-build] [the-bomb] [the-atomic] . ← gloss

wyntj AlmfAEI Alblwtwnywm AllAzm ItSnyE Alqnbip Al*ryp . ← Romanized text

w+ yntj AlmfAEI Alblwtwnywm AllAzm l+ tSnyE Alqnbip Al*ryp . ← MADA Morphological Analyzer

- ▶ longer sentences, but fewer distinct tokens

Sources of variation

- ▶ Polysynthetic languages – *Siberian*
A single word might correspond to a single English sentence
- ▶ Agglutinative languages – *Turkish*
Morphemes are segmentable
- ▶ Fusion languages – *Russian*
affixes and prefixes carry syntactic meaning

Movement – Variations in Word Order

Languages are often classified by the ordering of their subject, verb, and object clauses .

- ▶ S-V-O – *English*
- ▶ S-O-V – *Japanese*
- ▶ V-S-O – *Arabic*

For example, in translating from Arabic into Japanese, the verb might have to be ‘moved’ from a sentence initial position to a sentence final position.

Movement leads to computational difficulties in machine translation

- ▶ Ordering changes in translation are not deterministic
- ▶ The number of possible reorderings grows exponentially with sentence length
- ▶ Considering arbitrary reorderings in automatic translation can be an NP-Complete problem³

³K. Knight, "Decoding Complexity in Word-Replacement Translation Models," *Computational Linguistics*, 25(4), 1999

Names Are Often Transliterated Rather Than Translated

Transliteration is writing a word from one language using the *closest* corresponding orthography of a different language

- ▶ *Closeness* is often defined ‘phonetically’, i.e. the transliteration attempts to describe the pronunciation of a foreign word
- ▶ Transliteration of English into Chinese finds the syllables corresponding to a Chinese approximation to the English name ⁴
 - ▶ “Bush” → /bu shu/
 - ▶ “Clinton” → /kk lin dun/

- ▶ A written form with that pronunciation is then chosen

Transliterations are rarely unique and are often difficult to spot in “foreign” text

- ▶ As another example, “Kosovo” can be transliterated into Chinese in several ways

科索沃 /ke-suo-wo/, 科索佛 /ke-suo-fo/,
 科索夫 /ke-suo-fu/, 科索伏 /ke-suo-fu/, or
 柯索佛 /ke-suo-fo/.

and all of the above should be translated back into English as “Kosovo”

- ▶ Machine Translation systems must
 - ▶ Identify names that are to be transliterated rather than translated
 - ▶ Identify transliterations found in the foreign text and normalize them prior to translation

⁴Meng *et al.*, “Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval”, ASRU 2001

Statistical Machine Translation and Automatic Speech Recognition

Like ASR, SMT can be formulated using a Source-Channel model.



For a given English sentence E , the Translation Model $P(F|E)$

- ▶ provides the likelihood that any foreign sentence F is a valid translation
- ▶ 'generates' foreign sentences F probabilistically from 'E'

Both the transcription and the translation are found as Maximum A Posterior estimates.

Transcription

Input - an English utterance A

Output - an English transcription \hat{W}

$$\begin{aligned} \hat{W} &= \operatorname{argmax}_W P(W|A) \\ &= \operatorname{argmax}_W \frac{P(A|W) P(W)}{P(A)} \\ &= \operatorname{argmax}_W \underbrace{P(A|W)}_{\text{Acoustic Model}} \underbrace{P(W)}_{\text{Source Language Model}} \end{aligned}$$

Translation

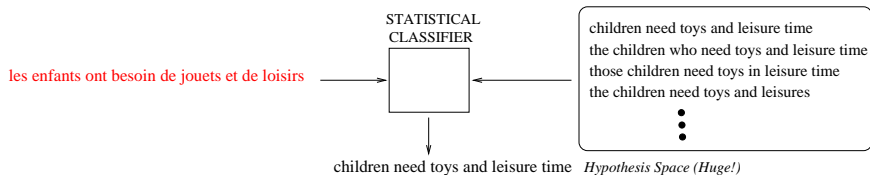
Input - a foreign sentence F

Output - an English sentence \hat{E}

$$\begin{aligned} \hat{E} &= \operatorname{argmax}_E P(E|F) \\ &= \operatorname{argmax}_E \frac{P(F|E) P(E)}{P(F)} \\ &= \operatorname{argmax}_E \underbrace{P(F|E)}_{\text{Translation Model}} \underbrace{P(E)}_{\text{Source Language Model}} \end{aligned}$$



Statistical Machine Translation and Pattern Recognition



Source-Channel Formulation: Input - a Foreign sentence F , Output - an English sentence \hat{E}

$$\hat{E} = \operatorname{argmax}_E P(E|F) = \operatorname{argmax}_E \frac{P(F|E) P(E)}{P(F)} = \operatorname{argmax}_E \underbrace{P(F|E)}_{\text{Translation Model}} \underbrace{P(E)}_{\text{Source Language Model}}$$

- ▶ The Source Language Model $P(E)$ can be estimated using *monolingual text* using the same techniques developed for use in speech recognition
- ▶ The next lecture will focus on the Translation Model $P(F|E)$
 - ▶ How should the model be formulated to capture the linguistic phenomena discussed earlier?
 - ▶ How can the model be formulated so that its parameters can be estimated from collections of translations, known as *Parallel Texts*?

Parallel Texts for Training and Evaluation of SMT Systems

A parallel text corpora consists of translations in two (or more) languages.

Parallel text can be **aligned** to identify **translation equivalence**

Translation (and therefore alignment) is a multi-level, hierarchical process

- ▶ Documents are aligned within collections ...
- ▶ Paragraphs are aligned with documents ...
- ▶ Sentences are aligned within paragraphs ...
- ▶ Words are aligned within sentences ...

Aligned elements in parallel texts are translations of each other.

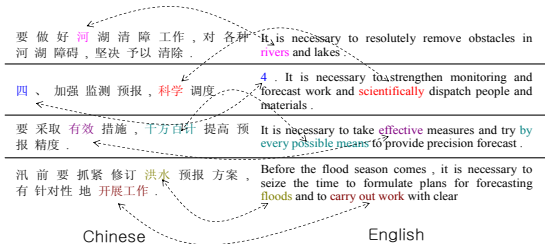
Alignments can be obtained in several ways

- ▶ Manually – Human translators mark alignment at the appropriate level
- ▶ Semiautomatically - Monolingual texts contain information that can be easily identified and used in alignment across languages, e.g. Chapter Headings, Verse Numbering, etc.
- ▶ Automatically – using statistical models, as will be discussed

Parallel text collections can be quite large – millions of words

- ▶ Parameters of statistical translation models can be estimated from these collections
- ▶ SMT systems can be **scored** against known translations

Chinese and English Text Aligned at the Sentence and Word Levels



Sentence Aligned French-English Parallel Text

Parallel documents with alignment information inserted :

[1] Perhaps the Commission or you could clarify a point for me. [2] It would appear that a speech made at the weekend by Mr Fischler indicates a change of his position. [3] I welcome this change because he has said that he will eat British beef [4] and that the ban was imposed specifically for economic and political reasons.

[1] La Commission ou vous-même pourriez peut-être m'expliquer un point. [2] Il semblerait en effet que M. Fischler ait changé de position dans un discours prononcé au cours de ce week-end. [3] Je me félicite de ce changement, car il a dit qu'il mangerait du boeuf britannique [4] et que l'interdiction avait été décrétée spécifiquement pour des raisons économiques et politiques.

- ▶ 'Markers' are inserted into the text to indicate the start and end of translation segments
- ▶ Translation segments can be sub-sentence units, e.g. at [4]