Технологията Преводна Памет

Елена Косева

Translation Memory Technology. *This paper proposes a brief overview of TM systems. System point of view regards TM systems as part of CMS or as a standalone CMS, namely GTMS intended to support the work on translation projects, and to conform to industry standards concerning the exchange of content. From computational linguistics' point of view, TM systems are applied field of computational linguistics concerning segmenting, parsing, storage and searching models, aligning and measuring distance between strings.*

Key words: *translation memory, CMS, string searching, bilingual text, distance measuring, computational linguistics*

ВЪВЕДЕНИЕ

Автоматичният превод, известен още като машнен превод, а в още по-ранните години – механичен превод, е проблем, по който се работи повече от 60 години. Независимо от големия брой изследвания и прилагането на различни модели и подходи, наличните системи и технологии не са в състояние да превеждат висококачествено и напълно автоматично естественоезикови текстове.

Още през 1966 г. комисията ALPAC (Automatic Language Processing Advisory Committee) излиза със становището, че висококачествен напълно автоматичен превод е невъзможен и дава препоръка да се насочат усилията на разработчиците в посока на приложения, подпомагащи работата с лингвистични данни както и към създаването на полезни лингвистични ресурси [1]. Продукт на такъв подход са системите за компютърно подпомогнат превод (CAT Computer Aided Translation), наричани още системи за автоматизиран превод. Същата комисия предлага и и термина Компютърна Лингвистика (Computational Linguistics) като събирателен термин за всички езикови приложения, изпълнявани на компютър и прилагащи различни нива на анализ и обработка предимно на естествени езици в писмена или речева форма.

Системите използващи преводна памет (TMS Translation Memory Systems) са софтуеърни приложения, които са особено актуални в днешно време поради огромния брой документи от различен тип подлежащи на превод, в това число и за превеждане на документацията и помощните файлове на софтуеърни приложения, от които се изисква интернационализация и локализация.

През 1970 г. се появява първото приложение използващо принципа на днешните преводни памети. Наричало се е Repetition Processing и е доставяло извадки от преведени вече текстове като използваният критерий е бил близостта (сиmиларити) с текущо превеждания текст. През 80^{те} години на XX век са добавени приложение за редактиране, приложение за управление на терминологията и приложение за автоматично търсене в речник, с което структурата на преводните памети почти съвпада с тази на сега съществуващите. По-късно са предложени подобрения в посока търсене и използване на подобни сегменти в текущо превеждания текст. Като цяло идеята за преводна памет се е появила в средите на разработчиците на софтуеърни приложения. Първите комерсиални решения се появяват в средата на 90^{те} години на миналия век.

Въпреки че наличните решения се различат в конкретните си реализации, основна отличителна характеристика е моделът на паметта:

1. Модел база данни или сегментен подход най-често на ниво изречение (Sentence-Based-Approach);
2. Модел битекст – (Character-String-in-Bitext CBS).

Дефиницията за преводаческа памет дадена от EAGLES (Expert Advisory Group on Language Engineering Standards) е: „Многоезичен текстов архив, съдържащ сегментирани, изравнени, тагирани като части на речта или синтактично анотирани

и класифицирани многоезични текстове, позволяващ съхранение, търсене и доставяне на изравнени многоезични текстови сегменти, отговарящи на различни критерии за търсене."

Системите използващи преводана памет могат да бъдат разгледани от две гледни точки:

1. Като самостоятелни модули в Системите за управление на съдържание (CMS Content Management Systems), предоставящи функционалност, подпомагаща превода или локализацията на контента, а също и като напълно самостоятелни системи Global Translation Management Systems [3];
2. Като системи прилагащи постиженията на компютърната лингвистика и езиковото инженерство т.е. като приложения за обработка на езици.

ИЗЛОЖЕНИЕ

Терминът Система за Уравление на Съдържание[2] става популярен най-вече чрез системите за управление на уеб съдържание и по-точно системите за управление на динамични уеб сайтове (WCMS[3]). Съществуват различни видове системи за управление на съдържание: Системи за управление на документи и записи (DMS и RMS), Системи за управление и организиране на учебния процес (Learning Management System), Системи за управление на библиотеки (Library Management System).

Системите с преводна памет[4] включват в състава си типичните за всяка CMS инструменти за настройка на работната среда, инструменти за управление на работния поток (workflow), средства за поддържане на версии, средства за управление на жизнения цикъл на контента, възможности за редактиране (добавяне и изтриване) на преводната памет, управление на терминология, възможности за обмен на съдържание с други системи. Според спецификата на контента всяка CMS използва и различни допълнителни инструменти.

Когато е необходим превод на предлаганото от CMS съдържание, изборът на средство за превод зависи до голяма степен от възможността за предварителен контрол при изготвяне на контента. Когато е възможно от гледна точка на контента, решение може да бъде дори използване на напълно автоматичен превод, при условие, че са налични инструменти за създаване на контент, в които се използват контролиран език и контролиран речник т.е. наложени са синтактични, структурни и терминологични ограничения.

Ако е невъзможно ограничаване на лексиката, по-удачното решение е използване на автоматизиран превод – преводна памет. От дефиницията на преводна памет е ясно, че използването й е удобно когато се очаква висок процент повторяемост на използваните термини. Това е възможно за някои технически текстовете, и за такива, които подлежат често на малки промени т.е. създават се различни версии. На тези критерии отговаря също придружаващата документация към изделия предназначени за глобалния пазар, различни ръководства за употреба, закони, правилници и наредби, медицински текстове от типа указания за употреба на лекарствени средства и др. Когато се създават обновени версии на вече съществуващи документи степента на повторяемост може да бъде много висока като се случва да се повтарят не само изрази и изречения, а дори цели параграфи. Различните източници посочват проценти на съвпадение между съществуващите вече документи и новите им версии между 20% и 70%.

Като самостоятелни системи за управление на съдържание системите с преводна памет са интересни от гледна точка на архитектурата и съвместимостта

[2] Терминът е контент, но се употребява и съдържание.

[3] Web Content Management System;

[4] Преводните памети имат и други наименования, най-популярното от които е банка с данни – data bank

със стандартите, отнасящи се до естеството на обработвания контент. Типичната система с преводна памет, се състои от набор от инструменти, някои от които са:

- Инструменти за сегментиране, с тях се променя размерът на преводната единица. Стандартът за обмен на съдържание на преводни памети приема по подразбиране преводна единица изречение, но в някои случаи е възможно по-удачна (от гледна точка на процент съвпадения) да се окаже работата с по-малка преводна единица като дума или сегмент[5]. Това е инструмент, който е основен в системи с модел база данни и се използва за съвместимост със стандарта TMX в системи с модел битекст.
- Инструменти за управление на терминология. Обикновено се наричат терминологична база данни[6]. В най-простия си вид терминологичната база данни се състои от преводите на термините специфични за конкретна предметна област на поддържаните от инструмента езици. В по-разширените варианти освен термините в базата данни се съхранява информация за граматичната категория на термина, примерна употреба в изречение, информация за контекстна зависимост и др. Обикновено търсенето и замяната на термини от терминологичната база данни работи във фонов режим.
- Преводна памет с модел базата данни съхранява изрази, сегменти, изречения или фрази в зависимост от избраното ниво на сегментация. При модел битекст, преводната памет представлява множество от файлове с текстове и файлове със съответните им преводи, които се намират достъпни в работното пространство и могат да бъдат селектирани по различни критерии чрез съпътстващите ги метаданни. Има и някои изпълнения, при които битекстовите файлове се намират извън работното пространство и се свързват към приложението. Този вид организация се нарича виртуална преводна памет.
- Инструмент за работа на преводача – текстов прозорец, в който по подходящ начин се индицират намерените в преводната памет съвпадения. Съвпаденията най-общо са пълни (exact match) и непълни или приблизителни (fuzzy match). В общия случай и двата вида съвпадение се предлагат на потребителя за одобрение. Тук интересен е въпросът за определянето на прага на процентното съвпадение при непълно съвпадение. Този праг е езиково зависим и е настройваем параметър. За английски текстове преводачите го определят около 90%.
- С цел повишаване на ефективността и осигуряване на информация за настройка на системата са налични и инструменти за събиране на статистически данни от типа процент съвпадения, процент приети от преводача пълни или приблизителни съвпадения и др.

От гледна точка на поддържани стандарти. Необходима е съвместимост със стандартите за обмен на съответния тип контент, а именно: TBX (Term Base eXchange), TMX (Translation Memory eXchange), SRX (Segmentation Rules eXchange) и XLIFF (XML Localization Interchange File Format).

TBX е отворен стандарт за обмен на структурирани терминологични данни. Това е стандарт създаден от LISA Localization Industry Standards Association OSCAR (Open Standards for Container/content Allowing Reuse) и възприет от ISO като ISO 30042: 2008 Systems to manage terminology, knowledge and content – Term Base eXchange (TBX) [7].

TMX е отворен XML стандарт за обмен на данни от преводни памети създадени от CAT инструменти и инструменти за локализация. Целта на TMX е да улесни процеса на обмен на данни от преводни памети между различните инструменти и

[5] сегмент е българският термин за фраза, която не съдържа подфраза от собствения си вид (chunk)

[6] term bank

производители на преводачески инструменти с малка или никаква загуба на ценни данни по време на трансфера.

SRX е стандарт за описание на начина, по който инструментите за превод или други инструменти за езикова обработка сегментират текст. С негова помощ инструменти за преводни памети и други лингвистични инструменти могат да описват езиковоспецифичния процес, при който текстът се разделя на сегменти с цел понататъшната им обработка. Стандартът е бил разработен когато е станало ясно, че ползата от TMX е понякога по-малка от очакваната поради това, че различните инструменти сегментират текста по различен начин. Когато се прилага заедно с TMX, SRX осигурява предаването на правилата за сегментация, които са били приложени в процеса на създаването на преводната памет така, че да се подобри използваемостта на обменените данни. SRX може също да се използва от всеки инструмент, който сегментира текст с цел подобряване на интеграцията с други процеси [8].

Работата на XLIFF е базирана на концепцията извличане на контента от източника, който трябва да бъде преведен или локализиран и след това връщането му отново на съответното място. Използваният модел се нарича Extract/Merge (изваждане и сливане). Начините за отделяне на подлежащия на превод контент от форматиращите, маркъп данните или други неподлежащи на превод данни са два: капсулиране и метод с контейнери съответно за съвместимост с TMX и OpenTag. Освен това стандартът осигурява схема за индициране на кандидатите за превод – това са намерените пълни или частични съвпадения с процент на съвпадение по-голям от определения праг [8].

От гледна точка на компютърната лингвистика, преводните памети са един вид система за автоматизиран превод. Предимството на тази система пред напълно автоматичния превод е доброто отношение качество/време за превод, което се дължи на това, че някои от основните проблеми на машинния превод: синтактичната и семантична многозначност и неопределеност както и изборът на подходящата дума или термин съобразно регистъра на дискурса се разрешават от човек преводач, който при това е висококвалифициран. Двата модела на организация на преводните памети използват различни алгоритми за търсене [4].

При модел база данни, използваният алгоритъм оценява растоянието между сравняваните стрингове в термините на брой изтривания, замествания или вмъквания т.е. разстоянието е edit distance. Най-често това е разстояние по Левинщайн. При този модел предположението, неформално казано, е че всяко изречение има само един допустим превод. Така е само от гледна точка на модела, а като реализиция има някои системи, които използват парафразиране като начин за подобряване на процента съвпадения. Върнатият резултат може да е пълно съвпадение (exact match) или частично съвпадение (fuzzy match). При този модел настройваем параметър е и преводната единица, която е равна на сегмента.

Вторият модел използва алгоритъм, който търси в несегментиран текст най-дългите съвпадащи стрингове. При това съвпаденята могат да са само пълни и не са ограничени по дължина. Те могат да излизат извън границите на изречение, дори параграф. Търсят се всички съвпадения с произволна дължина в набор от избрани по някакъв критерий двойки съответни документи. Това, което получава преводачът като резултат е подадения текст, в който са заместени всички намерени съвпадения т.е. смес от преведени и непреведени стрингове. Този модел прилича на статистическия модел на машинния превод по това, че използва изравнени битекстове, които са вид корпуси. Битекстовете се събират от преводача в процеса на работата му. За да са годни за употреба се изравняват на ниво изречение и се съхраняват като преди това се маркират с подходящи метаданни. Те се използват при избиране на релевантните документи за текущия проект. Метаданните са от типа: автор, тематика, дата на създаване, проект за който е създен битекстът и др.

ЗАКЛЮЧЕНИЕ

Всички по-известни ситеми, независимо от използвания модел на паметта, са в състояние да експортират паметите си в TMX формат. Изборът на конкретен модел на паметта се определя най-вече от типа на превежданите текстове, по–точно от това съвпадения на стрингове с какви дължини се очакват. Преобладаващата част от наличните в момента системи, независимо от типа си, комерсиални или свободни и с отворен код, използват модел на паметта база данни.

Доколкото концепцията на технологията преводни памети е увеличаване на ефективността чрез търсене на преведени вече текстове, посоката в която може да се търси още по-голяма ефективност, е повишаване процента на пълните съвпадения и по-голям процент на съвпаденията при непълните съвпадения. Направените подобрения в последните версии на известните системи акцентират върху използване на парафразиране (nested match), оценяване на съвпадението в контекст (perfect match), използване на машинен превод, генериране на предположения (suggestions). Други насоки, в които може да се търси са видоизменения на алгоритмите за оценяване на съвпаденията, използване на синонимни множества, за някои езици може да се окаже подходящо използване на списъци с фразови глаголи или търсене на двойки думи например предлог и думата след него. В това отношение в литературата се посочват опити с алгоритми за сравнение и оценяване на близостта на ниво дума както и използване на статистически методи за превод – превод базиран на примери (EBMT Example Based Machine Translation) [5].

ЛИТЕРАТУРА

[1] Иванов. И., Й. Захариева, П. Стойков. Комютърната лингвистика като съставен елемент на изкуствения интелект. Фараго, София, 2008.

[2] Николов. Л., С. Бонев. Формални езици и езикови процесори, Издателство на ТУ – София, София, 2005.

[3] Alison Toon, Andrew Draheim, Arle Lommel, Pierre Cadieux, Edited by Rebecca Ray. LISA Best Practice Guide: Managing Global Content. Global Content Management and Global Translation Management Systems, LISA, 2007.

[4] Elina Lagoudaki. Translation Memory Systems: Enlightening users' perspective, Imperial College London, London, 2006.

[5] Francie Gow, Metrics for Evaluating Translation Memory Software. Thesis submitted to the Faculty of Graduated and Postdoctoral Studies of the University of Ottawa, < http://www.chandos.ca/thesis.html> to date

[6] R. Asanka Wasala. Initial Survey on the Availability of Translation Memory Tools, PAN Localization Project, University of Colombo School of Computing, Sri Lanka, 2008.

[7] Translation Memory eXchange, Term Base eXchange, Segmentation Rules eXchange < http://www.lisa.org/ > to date

[8] A white paper on version 1.2 of the XML Localisation Interchange File Format (XLIFF), <http://www.oasis-open.org/committees/download.php/26817/xliff-core-whitepaper-1.2-cs.pdf> to date

[9] Linux for translators <http://www.linuxfortranslators.org/> to date

За контакти:

Елена Косева Косева, УНИБИТ - София, катедра „Информационни технологии", редовен докторант Тел. 02 9876334, GSM 0898 258 795, e-mail: peblestel@gmail.com

Докладът е рецензиран.