УНИВЕРСИТЕТ ПО БИБЛИОТЕКОЗНАНИЕ И ИНФОРМАЦИОННИ ТЕХНОЛОГИИ

КУРСОВА РАБОТА
по
Компютърна лингвистика

Тема: *„Машинен превод“*

Изготвил:

I курс, магистърска програма „Информационни технологии“

гр. СОФИЯ
20.05.2015 г.

Ръководител: ……..

Съдържание

I. Същност и история на машинния превод

1.1. Същност

Машинният превод е дисциплина от компютърната лингвистика, която се занимава с автоматичен превод на писмен текст или реч от един естествен език на друг, с помощта на компютърен софтуер. Най-простият вариант на машинен превод е заместването на една дума с друга. За постигане на по-сложни методи за превод се използват текстови корпуси. Те дават възможност за отразяване на типологичните различия между езиците, разпознаване на фрази, превеждане на идиоми и изолиране на аномалии.

Съвременните софтуери за машинен превод позволяват избора на предметна област. По този начин се подобрява качеството на превода чрез ограничаване на допустимите значения на думите. Това е и причината административните и правните текстови да имат по-сполучлив превод от разговорната реч.

През 1949г. Уорън Уейвър от фондация Рокфелер пише „*Пред мен стои текст, написан на руски, но аз ще си мисля, че всъщност е на английски, но е кодиран с някакви странни символи. Всичко, което трябва да направя, за да разчета закодираната информация, е да го декодирам*.” Макар привидно наивно, това разсъждение се оказва изключително полезно и днес лежи в основата на съвременния статистически машинен превод[1]. Затова и процесът на превеждане се

[1] http://people.ischool.berkeley.edu/~nakov/selected_papers_list/nakov_prevod_sp_Avtomatika_Informatika.pdf

разглежда като съвкупност от декодиране на значението на входящия текст и кодиране на значението със средствата на естествения език.

1.2. История

Развитието на машинния превод започва през 50-те години на XX век, по време на „студената война“. През 1954г. е проведен експеримент „Джорджтаун - IBM“, с участието на университета Джорджтаун и фирмата IBM, който демонстрира превода от руски текст на английски. Експеримента отбелязва голям успех и авторите му предсказват бърз и голям напредък на автоматичния превод. Действителния напредък се оказва доста по-бавен.

България също не изостава от тенденциите и през 1964г. се създава специална група за машинен превод между руски и български език под ръководството на проф. Александър Людсканов в Института по математика на БАН.

През 1966г. настъпва обрат. По поръчка на правителството на САЩ, специален комитет по приложна лингвистика (ALPAC) към Националната академия на науките на САЩ изготвя доклад за състоянието на изследванията за машинния превод, който се оказва силно скептичен. В резултат на печално известния доклад „Черната книга на машинния превод“, настъпва етап на застой на изследванията.

Едва през 1975-1985г. започва постепенно възраждане първо в Европа, Япония и СССР, а след 1985г. – и в САЩ. Истински обрат настъпва в края на 80-те години на XX век, когато компютрите набират голяма изчислителна мощ и в същото време поевтиняват и стават по-достъпни и възраждат интереса към машинния превод.

II. Трудности пред машинния превод

Превода се характеризира с гладко предаване на значението и смисъла от един език на друг. За човек това не представлява особена трудност, тъй като той знае поне единия език. Когато става въпрос за машинен превод обаче, компютъра не разбира нито един от двата езика. Затова той разчита на модели и правила, с които да може да осъществи превода.

Всеки език обаче има своите особености и често различията между два езика са доста по големи от просто различно изписване на думите. Междуезиковите различия може да се разгледат на няколко нива:

2.1 Словоред

В повечето европейски езици словореда е подлог-сказуемо-допълнение :

например англ. "I like beer." („Аз обичам бира.")[2]

В други езици като турски, японски и хинди той е подлог-допълнение-сказуемо :

например тур. „Ben bira seviyorum"

(буквално „Аз бира обичам.")[2]

Езици с падежни форми като българския и чешкия имат относително свободен словоред и позволяват спокойно поставяне на

[2] Посочените примери са от : http://people.ischool.berkeley.edu/~nakov/selected_papers_list/nakov_prevod_sp_Avtomatika_Informatika.pdf

думите на различни позиции. Така на пример посоченото изречение по-горе на български може да изглежда по следните начини:

„Аз обичам бира.", „Аз бира обичам.", „Бира аз обичам.", „Бира обичам аз.", „Обичам аз бира.", „Обичам бира аз."[3]

Разликите в словореда имат и други характеристики. В немския език спрегнатия глагол винаги трябва да е на второ място, а подлога и прякото допълнение могат да си разменят местата.

2.2 Отрицанието

И тук повечето европейски езици си приличат. В голяма част от тях отрицанието се изразява със специфична частица, която изразява отрицание, и стой непосредствено преди глагола. В немския обаче отрицанието стой в края на изречението, като има и изключения в сложните времена – тогава отрицанието се поставя непосредствено преди глагола:

бълг. „<u>Не</u> искам бира."
исп. „<u>No</u> quiero cerveza."
итал. „<u>Non</u> voglio birra."
немски „Ich will Bier <u>nicht</u>.",
и сложно време Ich bin <u>nicht</u> spazieren gegangen[3]

Отрицанието може да се изразява и с различни думи в различните смислови ситуации. Също може да се ползват и няколко думи както е в английския, където често се изисква употребата на спомагателен глагол и др.

[3] Посочените примери са от : http://people.ischool.berkeley.edu/~nakov/selected_papers_list/nakov_prevod_sp_Avtomatika_Informatika.pdf

2.3 Род

В българския и на немския език има мъжки, женски и среден род. Има езици, в които липсва отличаване по род, като в турския и малайския например. Ангийския език пък прави разлика в родовете за одушевени предмети и в местоименията, но не и за предмети. Освен това рода на един предмет на един език може да бъде различен от рода на същия предмет на друг език. Разглеждайки рода в немския език, заради наличието и на падеж в този език, води до изменяне на начина на изписване на рода и може да породи проблеми в превода.

2.4 Число

На различните езици изразяването на граматическото число е различно и често източник на затруднения при превод. В повечето европейски езици, сред които и българския, има две форми – единствено число и множествено число. В словенски, руски и други славянски езици формата за множествено число при пет и повече предмета е различна от тази за два, три и четири.

2.5 Определителен и неопределителен член

Друг източник на трудност пред машинния превод е начина, по който в различните езици се поставя определителния и неопределителния член. В повечето балкански езици, той е част от думата, докато в по-голямата част от европейските езици той се поставя пред думата и представлява отделна част от изречението:

бълг. човекът / човека

нем. der Mann / ein Mann

анг. the man / a man

2.6 Лични местоимения

И тук има разнообразие в различните езици и още една спънка пред машинния превод. Така например в английския език, местоимението „ти“ (за неформално единствено число), „вие“ (за множествено число) и „Вие“ (форма на учтивост) са едни и същи (*you*) и няма как да се различат, докато например на италиански има четири форми.

2.7 Глаголи

Глаголните форми и времената са друга важна характеристика на езика. Тук разликите в различните времена на езиците може да доведат до разлики в смисъла на превода. Като пример може да се приведе изречение в сегашно време на български (*Аз уча английски*), което ако се преведе на английски ще трябва да се избира между *“I am studying English.” („Уча английски /в момента/.“)* и *“I study English.” („Уча английски /по принцип/.“)*. В езици като немския има и глаголи, които имат делими частици и те отиват в края на изречението, като целият смисъл на глагола зависи от тази частица.

2.8 Изпускане на подлога

За структурата на изречението на български не от значение дали подлога ще се изпусне, но в английския и немския например не може да се направи.

2.9 Структурни разлики

Има случаи, в които при превод се налага разместване на тематичните роли. Така например изречението „*Ти я харесваш“* ще има значими разлики на английски и на испански:

En: You like her. (Bg: Ти я харесваш.)

Sp: Ella te gusta. (Bg: Тя ти харесва.)[4]

Както е видно от примерите и буквалните им преводи на български и двата варианта са възможни в нашия език.

Също значими структурни разлики са използването на пряко допълнение с глагола в даден език, а за друг непряко; положението на предлозите и др.

2.10 Непреводими думи, реалии, лексикални дупки

По отношение на понятията и възприемането на света, често в различните езици има не само разлика в думите, но и за един предмет може да има няколко названия, а в друг език да е само едно или въобще да липсва. Така например английското *uncle,* в българския език съответстват няколко по-специфични думи: *вуйчо, чичо, свако, калеко, лелинчо*; на японски отсъства дума за брат, а вместо това има дума за голям брат и по-малък брат.[4]

2.11 Устойчиви словосъчетания, мултилексемни и идиоматични изрази и други многозначности

Трудностите тук са свързани с това, че може да има една и съща дума, която да се превежда различно от контекста. А на друг език при превод може да има дума, която да се изпуска.

Друго затруднение пред превода са идиомите. Те не могат да се превеждат буквално и значението им изискват познание на културата.

[4] Посочените примери са от : http://people.ischool.berkeley.edu/~nakov/selected_papers_list/nakov_prevod_sp_Avtomatika_Informatika.pdf

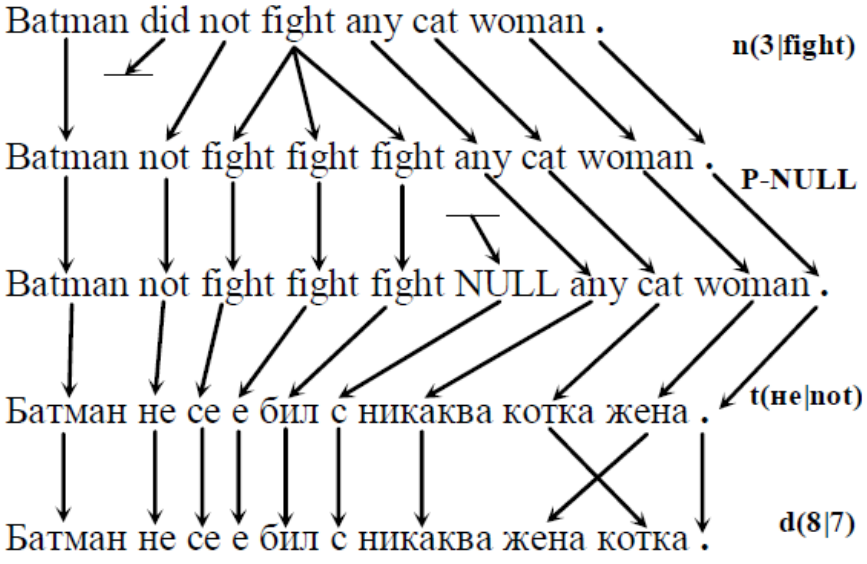
III. Подходи и модели за машинен превод

3.1 Моделът дума по дума

Най-простият вариант на машинен превод е дума по дума - без да се коригира словореда, граматиката или да се съобразяват различните значения.

Основания на граматически правила превод се учи от двуезичен паралелен корпус. Възможно решение в този вид превод дава статистически машинен превод, разработени от IBM през 1991г. В статия от 1993г. са публикувани 5 модела, станали известни по-късно като *модели на IBM 1, 2, 3, 4 и 5*.

Модел 3 на IBM е генеративен модел, който описва превеждането на изречение от английски на български в 4 стъпки, основани на съответните вероятности (*Фиг. 1*).



Фиг. 1[5]

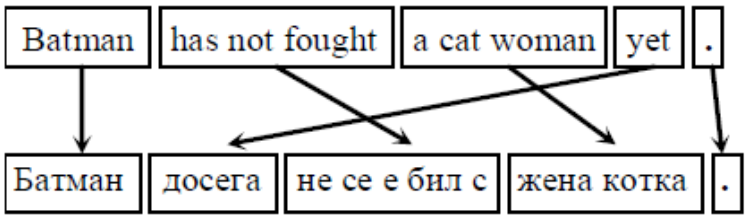
[5] Посочения пример е от : https://softuni.bg/trainings/1030/Statistical-Machine-Translation

Първата стъпка е да се решава, с колко думи ще се превежда всяка една английска дума. Втората, на някои позиции се вмъква празна дума. Третата, всяка английска дума се превежда със съответна българска. Накрая, някои от българските думи се разместват. С всяка от стъпките е асоциирана съответната вероятност (показана вдясно).

Това, което се сочи като проблеми на модела на IBM, е ограничение да учи отношения. Справя се с едно към много, но не може да се справи с отношения много към едно или много към много. Или иначе казано, базирайки се на превода, една английска дума може да сочи към много думи на български, но няма да се справи ако една българска дума може да се преведе с много английски. Друг проблем пред този метод е невъзможността да отчита контекстовото значение. Ако една дума означава различни неща в различни изречения, това няма да бъде отчетено.

3.2 Моделът превод с цели фрази

Този модел решава голяма част на проблемите на превода дума по дума. Той пак е генеративен модел и описва преобразуването на изречение от английски на български в стъпки. В този случай английското изречение се разбива на фрази и те се превеждат със съответстващите фрази на български. (*Фиг. 2*)

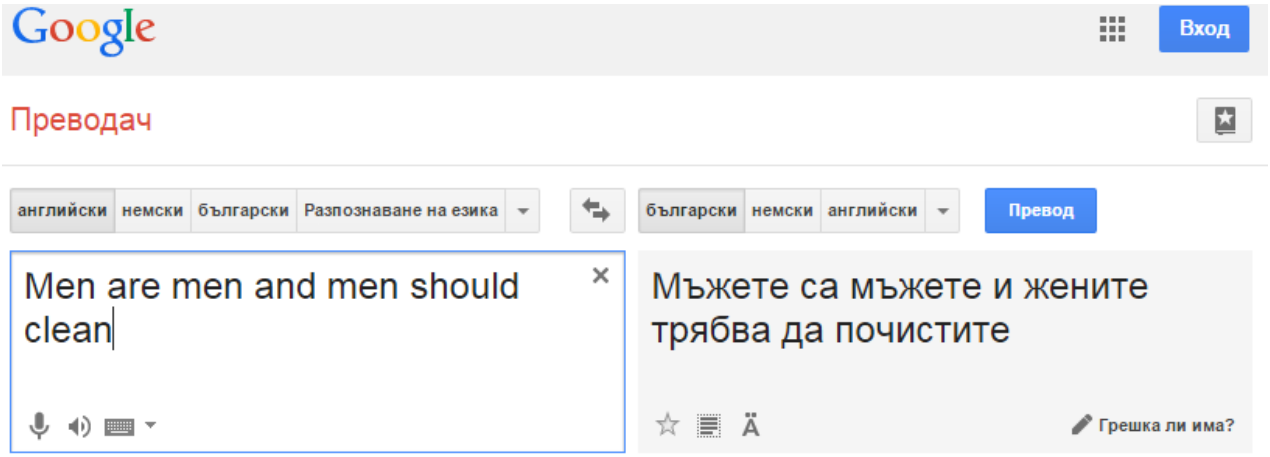


Фиг2[6]

[6] Посочения пример е от : https://softuni.bg/trainings/1030/Statistical-Machine-Translation

Този метод води до значително подобрение в машинния превод. При него вече е налице възможността да се учи отношение много към много. Също така се подобрява превода на контекстово значение. Освен това фразите могат да съдържат и препинателни знаци, което позволява да се научат разлики в правилата за пунктуация между двата езика.

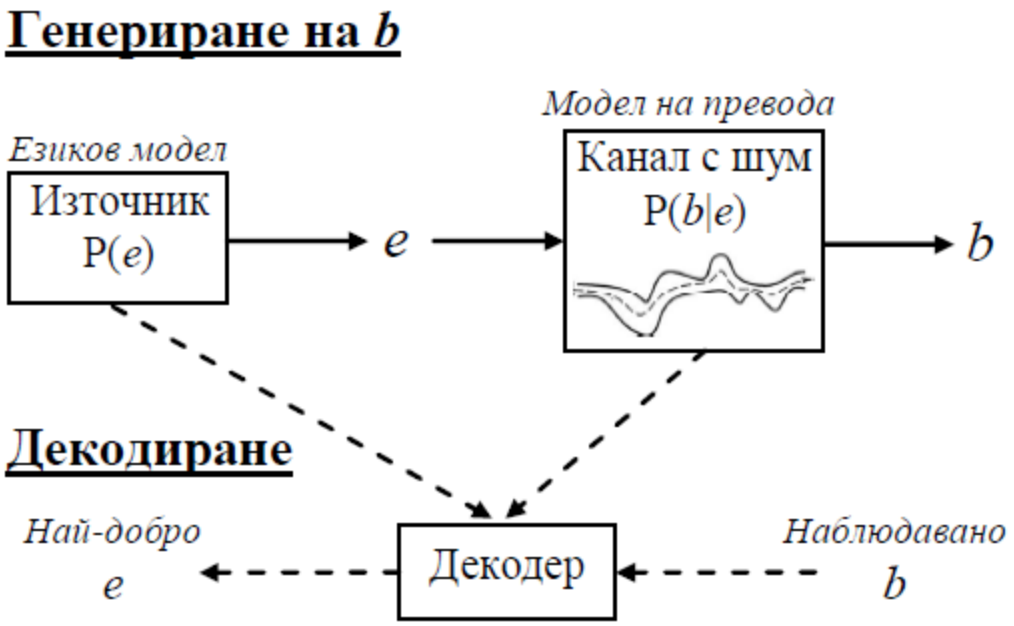
Тук също така може да възникнат трудности свързани с границите на фразите. Освен това този метод също е статистически. Той разчита на това колко често е срещана дадена фраза, за да предпочете нея за превод. Може да се стигне в предпочитание на дадена фраза и разминаване на значението и в превода (*Фиг. 3*).



Фиг. 3

3.3 Статистически машинен превод

В този модел данните за съответствието между думите и поредицата от думи в даден език се събира автоматично от така наречените двуезични корпуси. Процеса може да се опише чрез *предаване на информация по канал с шум* (*Фиг 4*).



Фиг. 4[7]

Съгласно този модел, българското изречение се разглежда като повреден вариант на английския оригинал. Моделът обяснява процеса на превода в две стъпки: първо се генерира английско изречение съгласно модела на източника, а след това то се предава по канала със шум и се поврежда съгласно модела на канала.

Разглеждането на този процес и възможните уравнения за решаване на вероятностите дават трите основни компоненти на една система за статистически машинен превод:

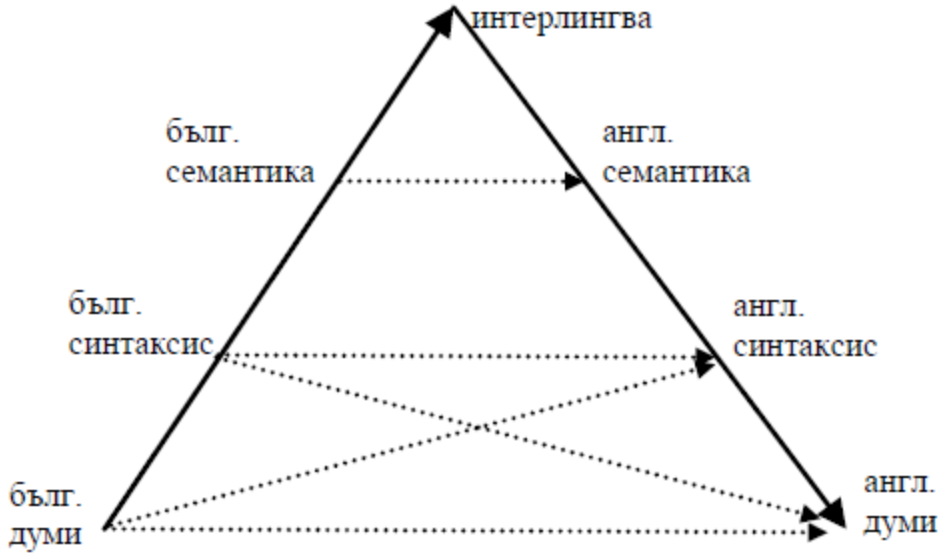
- езиков модел - показва колко е вероятно да бъде казано дадено английско изречение, като трябва да дава по-голяма вероятност на граматически правилното изречение за сметка на неправилното;

[7] Посочения пример е от : https://softuni.bg/trainings/1030/Statistical-Machine-Translation

- моделът на превода – интересува се единствено от това дали е вероятно двете изречения да са превод едно на друго, без значение дали английското (входното изречение) е граматично правилно построено;
- декодер - е търсещият алгоритъм, който по зададено изречение на даден език търси най-доброто съответствие за превод.

3.4 Машинен превод чрез междинен език

При този подход входящия текст се трансформира във вид, независим от входящия и от крайния език, наречен *интерлингва* (език посредник). Обособени са три нива на трансфер и интерлингва показани на *Фиг. 5*[8]*:*



Фиг. 5

[8] Посочената схема е от :
http://people.ischool.berkeley.edu/~nakov/selected_papers_list/nakov_prevod_sp_Avtomatika_Informa

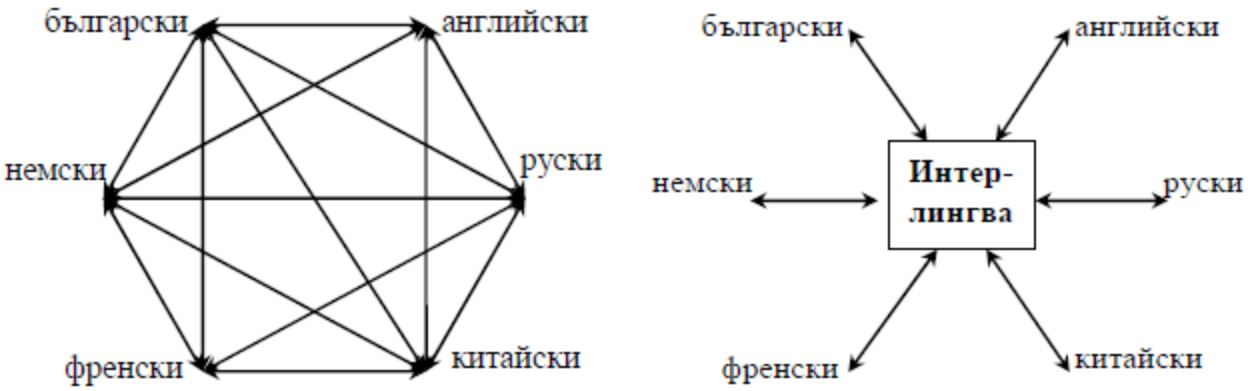
На схемата е показан идеалния случай, през който трябва да минава един превод чрез интерлингва. Като стъпки трябва да се премине през анализира на ниво дума, а след това и на синтактично и семантично ниво, преминава се в строеж на интерлингва, от която се генерира съответно английско семантично и синтактично представяне, и накрая – съответния английски превод.

В съвременните системи за машинен превод този процес обаче се съкращава. Както е показано на *Фиг. 5* това става по пунктираните линии, като най-често се ползват най-долните нива. Като развитието на машинния превод показва, че колкото по-нагоре се стига по стъпките посочени на триъгълника, това води до по-добри резултати в превода.

Създаването и употребата на интерлингва се стреми да улесни двупосочните процеси в преводите между различни езици. Като пример може да разгледаме връзките между 6 езика, които за да се преведат трябва да има 30 системи за машинен превод. С помощта на интерлингва се намаляват на 12 (*Фиг.6*)[9]. Дефинирането на подходяща интерлингва обаче засега остава нерешена задача.

tika.pdf

[9] Посочената схема е от : http://people.ischool.berkeley.edu/~nakov/selected_papers_list/nakov_prevod_sp_Avtomatika_Informatika.pdf


български
английски
немски
руски
френски
китайски
български
английски
немски
Интер-
лингва
руски
френски
китайски

Фиг. 6

IV. Заключение

Машинния превод е реалност. Той бележи напредък макар и с не бързите темпове, които първоначално са били очаквани. Въпреки че качеството на превод, който предоставя, е все още далеч от професионалния преводач, машинния превод вече се използва, намирайки своето приложение най-вече в уеб. Не случайно една от най-използваните функционалности на Google е функцията за автоматичен превод.

Развитието на машинния превод предстои. За подобряване на качеството не е достатъчен само статистически метод – да се гледат думите. Нужни са нови модели, обръщане внимание на словосъчетанията и семантичната релация. Напредъка, който Google и Skype не отдавна отбелязаха във гласовото разпознаване и превода на реч вдъхва надежди това да се случи и с автоматичния превод.

Това съвсем не означава, че преводачите ще станат излишни. Те вече използват така наречените *преводачески памети* (у нас е най-популярна е Trados), с което ускоряват процеса на превеждане. В САЩ вече има компании специализирани в редактирането на машинен превод, с което предоставят качествена и бърза услуга на конкурентни цени.

След появата на статистическия подход през 1991г. за превод, революция в това отношение се отбелязва през 2003г., когато е предложен ефективен модел за превеждане с цели фрази. И имайки предвид това и развитието, което непрекъснато се наблюдава в технологиите и разработките в сферата на изкуствения интелект оставаме с очакване на следващата революция.

V. Използвана литература

1. Стойков, П. - Компютърната лингвистика като елемент на изкуствения интелект – Фараго, 2012 г.
2. Наков, П. - Съвременен статистически машинен превод: кратък обзор http://people.ischool.berkeley.edu/~nakov/selected_papers_list/nakov_prevod_sp_Avtomatika_Informatika.pdf
3. softuni.bg - https://softuni.bg/trainings/1030/Statistical-Machine-Translation
4. blogs.skype.com – http://blogs.skype.com/2014/12/15/skype-translator-preview-an-exciting-journey-to-a-new-chapter-in-communication/
5. www.ibtimes.co.uk – http://www.ibtimes.co.uk/google-release-voice-recognition-language-translation-tool-1482968