Съвременен статистически машинен превод: кратък обзор

Преслав Наков,
Институт по паралелна обработка на информацията, БАН,
Секция по лингвистично моделиране
nakov@lml.bas.bg

Увод

Интересът към машинния превод – първата и една от най-важните задачи на изкуствения интелект – датира от 40-те години на миналия век и възниква още с появата на първите компютри. Това е периодът непосредствено след края на Втората световна война, по време на която в САЩ, във връзка с необходимостта от декодиране на съдържанието на прихванатите немски съобщения, силно развитие получават теорията на информацията и криптографията. Създаденият в резултат мощен математическо-статистически апарат съвсем естествено се разглежда като средство за постигане на автоматичен превод. През 1949 г. Уорън Уейвър от фондация Рокфелер пише [*Weaver,1955*]: „*Пред мен има 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 кодираната информация, е да декодирам текста.*" Макар привидно наивно, това разсъждение се оказва изключително полезно и днес лежи в основата на съвременния статистически машинен превод.

Идеята за автоматичен превод бързо буди ентусиазъм. През 1954 г. в САЩ се създава първият руско-английски прототип, което подсилва очакванията за бърз напредък, включително и за други двойки езици. Изследванията в България също не закъсняват и през 1964 г. се създава специална група за машинен превод между руски и български език под ръководството на проф. Александър Людсканов в Института по математика на БАН. Междувременно, във връзка със Студената война и започналото съревнование между СССР и САЩ в овладяването на Космоса, обемът на превежданата научно-техническа документация от руски на английски език нараства значително, а изследванията в областта на автоматичния превод се радват на богато финансиране. През 1966 г. обаче настъпва драматичен обрат: по поръчка на правителството на САЩ, Американската академия на науките изготвя доклад за състоянието на изследванията в областта на компютърната лингвистика и на машинния превод в частност, който се оказва силно скептичен [*Hutchins,2003*]. В резултат настъпва дълъг оздравителен период със силно ограничено финансиране първо в САЩ, а после и в световен мащаб. Едва през 1975-1985 г. започва постепенно възраждане.

Нещата рязко се променят през 90-те години на миналия век, когато компютрите набират достатъчно процесорна мощ, за да бъде проправен пътят на *статистическия* подход към машинния превод [*Brown&al.,1993*], който е доминиращ и до днес. Благодарение на последвалия значителен теоретичен и практически напредък, както и поради видимото подобрение в качеството, днес машинният превод отново се радва на значителен изследователски интерес, отчасти мотивиран от икономически очаквания – само годишните разходи за преводи на Европейската комисия възлизат на над един милиард евро годишно.

Защо машинният превод е труден

Постигането на качествен автоматичен машинен превод е една от най-важните, но и най-трудни, задачи на изкуствения интелект (ИИ). Счита се, че е „ИИ-пълна" (по аналогия с класа на NP-пълните задачи в теорията на алгоритмите), т.е. изисква решаване на всички фундаментални задачи на изкуствения интелект. От гледна точка на компютърната лингвистика качественият машинен превод предполага правилен автоматичен анализ и генериране на естествен език на морфологично, синтактично, семантично и прагматично ниво с отчитане на контекста, използване на знание за света и познаване на културата на носителите на езика. Изисква още генериране на текст, разрешаване на различни видове многозначност, разпознаване на собствени имена, транслитерация, анализ на местоимения, и изобщо дълбоко разбиране на смисъла на превеждания текст. Наред с това, много от предизвикателствата са чисто лингвистични: машинният превод, както и преводът изобщо, се усложняват заради обективните разлики между езиците.

Човешкият превод

Преди да видим как компютърът би могъл автоматично да превежда, например от български на английски език, нека първо помислим как го прави човек. Една възможна последователност на действията би могла да бъде следната:

1. *Четене* и *разбиране* на българския текст.
2. *Генериране* на съответен английски превод.

Съгласно това предположение, човек изгражда някакво междинно мислено представяне на смисъла (семантиката) на българския текст, от което след това генерира съответен превод. Дали при превод на друг език, напр. френски, това вътрешно представяне ще се промени или ще се запази (във втория случай имаме *интерлингва*)? Когнитивните специалисти и лингвистите не могат да дадат еднозначен отговор.

Всъщност, не е необходимо човек да е прочел и разбрал *целия* текст, за да започне да превежда (например при симултанен превод). Вместо това, обикновено преводът се извършва на части – изречения, малки фрагменти или дори отделни думи – като при нужда се взема предвид информацията в останалата част на текста.

Много учени отричат и необходимостта от дълбок анализ и цялостно разбиране при превод. Така например, през 1984 г. Макото Нагао дефинира следните два фундаментални принципа на превода [*Nagao,1984*]:

> *(1) Човек не превежда едно просто изречение чрез дълбок лингвистичен анализ.*
>
> *(2) Човек превежда, като първо разделя по подходящ начин изречението на фразови фрагменти, след което ги превежда с други фразови фрагменти, които накрая обединява по подходящ начин в цяло изречение. Преводът на всеки фразов фрагмент се извършва по принципа на аналогията, с използване на подходящи примери.*

Накрая, изглежда естествено, че за да може да превежда от български на английски език, човек трябва да владее и двата езика. Всъщност, това не е абсолютно необходимо: съгласно горните принципи, ако има достатъчно подходящи примерни преводи, които да използва за справка, човек би могъл да превежда по аналогия прости изречения и без да владее никой от двата езика. Така например, ако разполага със следните примери

Пример 1:

- англ: *(He buys) a notebook.*
- яп.: *(Kare ha) nouto (wo kau).*

Пример 2:

- англ: *I read (a book on international politics).*
- яп.: *Watashi ha (kokusaiseiji nitsuite kakareta hon) wo yomu.*

дори и да не владее нито японски, нито английски, човек лесно може да съобрази, че изречението

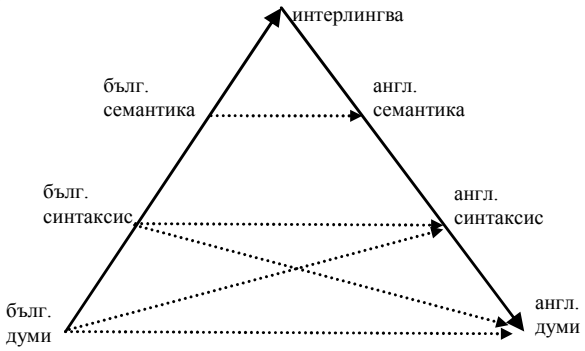
He buys *a book on international politics*.

би следвало да се превежда на японски като

(Kare ha) *(kokusaiseiji nitsuite kakareta hon)* (wo kau).

Така формулирана, задачата за машинния превод прилича на задача на математическата лингвистика, но е по-сложна, тъй като компютърът не владее нито един от двата езика, докато при математическата лингвистика човек най-често разсъждава върху съответствия между някакъв непознат език и своя роден.

Нива на трансфер и интерлингва

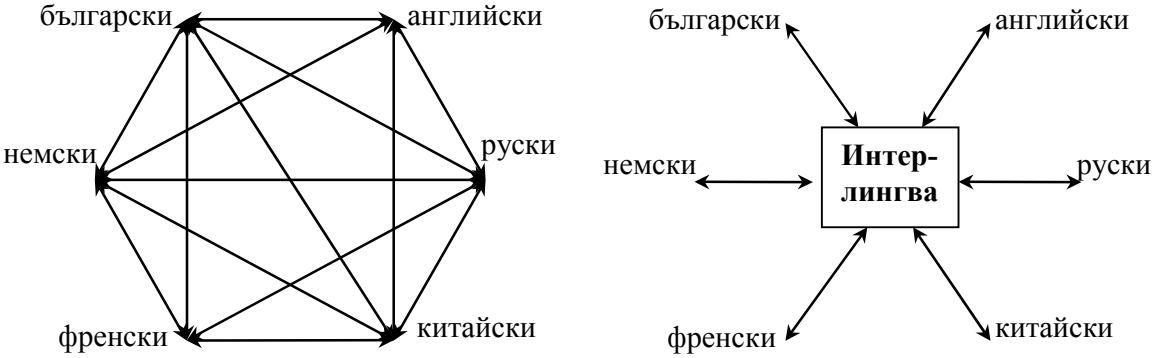


Фигура 1. Нива на трансфер.

Фигура 1 показва схематичен триъгълник с различни възможни нива на трансфер при превод от български на английски език. В идеалния случай българският текст първо се анализира на ниво дума, а след това и на синтактично и семантично ниво, след което се строи интерлингва, от която се генерира съответно английско семантично и синтактично представяне, и накрая – съответен английски превод. Повечето съвременни системи за машинен превод съкращават анализа по някоя от пунктираните линии, най-често от някое от най-долните две нива. Използването на семантично представяне – както езиково зависимо, така и езиково независимо (т.е. интерлингва) – засега остава извън възможностите на съвременните технологии за машинен превод, макар и да са правени някои опити в това отношение върху силно ограничени обеми данни.

Досегашното развитие на машинния превод недвусмислено показва, че всяко успешно изкачване нагоре по триъгълника от Фигура 1 води до по-добри резултати и до малка революция в машинния превод. За момента са овладени едва първите две нива, при това непълно: (1) превод на ниво думи – изцяло, и (2) използване на синтаксис – донякъде. Включително са овладени и някои междинни нива – превод с цели фрази и превод с йерархични фрази. Несъмнено стремежът към семантика и, в крайна сметка, интерлингва ще продължи, защото потенциалните ползи са несъмнени. Както показва Фигура 2, в об-

щия случай, за да се превежда в двете посоки между 6 езика, са необходими общо 30 системи за машинен превод, докато при използване на интерлингва са достатъчни 12, а при добавяне на седми език към системата ще са нужни съответно 12 и 2 допълнителни системи. Дефинирането на подходяща интерлингва обаче засега остава нерешена задача.



Фигура 2. Използване на интерлингва

Статистически машинен превод: модел на канала с шум

С цел простота на изложението по-долу ще предположим, че преводът се извършва изречение по изречение, без отчитане на съдържанието на останалата част от текста. Така работят повечето съвременни системи за машинен превод.

За по-голяма определеност ще предположим, че искаме да превеждаме от български на английски език. Така задачата за статистическия машинен превод може да се формулира по следния начин: *по дадено българско изречение b да се намери най-добрият му съответен английски превод e*. Ако означим с P($e|b$) вероятността *e* да бъде превод на *b*, получаваме следната формулировка: при дадено *b* да се намери *e*, за което вероятността P($e|b$) е максимална. Това ни води до уравнение (1), където с $\hat{e}$ сме означили търсеното максимално вероятно *e*. По правилото на Бейс, уравнение (1) може да се запише като (2), което от своя страна е еквивалентно на (3), тъй като знаменателят на (2) не зависи от *e*.

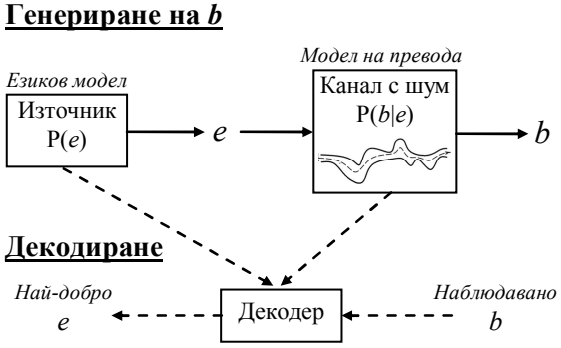
$$\hat{e} = \arg\max_e \mathrm{P}(e \mid b) \qquad (1)$$

$$= \arg\max_e \frac{\mathrm{P}(b \mid e)\mathrm{P}(e)}{\mathrm{P}(b)} \qquad (2)$$

$$= \arg\max_e \mathrm{P}(b \mid e)\mathrm{P}(e) \qquad (3)$$

Това е моделът на *предаване на информация по канал с шум*, приложен към статистическия машинен превод. Съгласно този модел, българското изречение *b* се разглежда като „повреден" вариант на английския оригинал *e*. Моделът обяснява процеса на генериране на българското изречение *b* като двустъпков процес: първо се генерира английско изречение съгласно *модела на източника* P(e), а след това то се предава по канала с шум и се поврежда съгласно *модела на канала* P($b|e$). Нашата задача е при зададени P(e), P($b|e$) и българско изречение *b* да намерим най-вероятното съответно английско изречение *e*. Моделът е показан графично на Фигура 3.

На пръв поглед не е ясно какво се постига с преобразуването на (1) в (3) – започнахме с P($e|b$), а в крайна сметка стигнахме до P($b|e$), което изисква превод в обратната посока, който е също толкова труден, както и в правата посока. В допълнение обаче получихме и допълнителен множител P(e), чието значение ще стане ясно по-долу.



Фигура 3. Преводът като декодиране.

Уравнение (3) всъщност ни дава трите основни компоненти на една система за статистически машинен превод:

- **езиков модел** P(*e*)
- **модел на превода** P(*b*|*e*)
- **декодер**, който по зададено *b* търси най-вероятното *e*

Езиковият модел P(*e*) показва колко е вероятно да бъде казано дадено английско изречение *e*. Моделът следва да дава висока вероятност на граматично правилни изречения (напр. „*I eat an apple*.") и ниска вероятност на неправилни (напр. „*Apple is eats I*."), както и на граматично правилни, но малко вероятни изречения (например в семантично отношение: „*The colourless green idea sleeps furiously*.").

Моделът на превода P(*b*|*e*) следва да дава висока вероятност за двойката (*b*,*e*), ако *b* би могло да бъде превод на *e*, и ниска вероятност – в противен случай. Забележете, че моделът на превода се интересува единствено от това дали е вероятно двете изречения да са превод едно на друго, без значение дали *e* е граматично правилно. Последното е задача на езиковия модел P(*e*). Разделянето на двете задачи в отделни модели позволява от една страна езиковият модел да се опрости значително, а от друга – създава възможност двата модела да се обучават поотделно, включително от различни изследователски групи. Възможно е дори обучаване върху различни данни: както ще видим по-долу, моделът на превода изисква паралелен двуезичен текст, който се събира трудно, докато езиковият модел използва само английски език, който може да се намери лесно в големи обеми.

Декодерът е търсещият алгоритъм, който по зададено *b* се опитва да намери най-вероятното *e*, т.е. това, което максимизира произведението P(*e*).P(*b*|*e*). Например, ако превеждаме *жената котка* на английски, това произведение ще бъде голямо за *the cat woman*, но малко за *woman cat the* (поради малък първи множител), за *Batman forever* (поради малък втори множител) и за *dog an a for* (заради двата множителя). Виж Таблица 1.

	P(e)	**P(f|e)**	**P(e).P(f|e)**
woman cat the	👎	👍	👎
woman the cat	👎	👍	👎
a cat woman	👍	👎	👎
cat woman	👍	👎	👎
the woman	👍	👎	👎
Batman forever	👍	👎	👎
dog an a for	👎	👎	👎
the cat woman	👍	👍	👍

Таблица 1. Как се превежда на английски език „жената котка"? *Добрият превод трябва да бъде едновременно граматично правилен и смислово верен.*

Моделът на канала с шум отразява факта, че човек превежда по-лесно от чужд език на своя собствен, отколкото в обратната посока: макар че „човешкият модел на превода" вероятно е еднакво добър в двете посоки, „човешкият езиков модел" е много по-добър за родния език, което прави преводите към този език по-гладки. Това се взема предвид от институции като Европейската комисия, която изисква официалните й документи да бъдат превеждани от преводачи, за които езикът, *към* който се превежда, е роден. Преводачи, за които роден е само езикът, *от* който се превежда, се считат за недостатъчно квалифицирани. По подобен начин при ръчно оценяване на система за машинен превод най-често се използват два отделни 5-степенни критерия: *адекватност* (каква част от информацията се съдържа в превода: цялата, повечето, много, малко, никаква) и *гладкост* (доколко гладък е изказът: безупречен, добър, "нероден", несвободен, неразбираем).

След 2002 г. доминиращият начин на оценяване е автоматичен и използва *BLEU* (*Bi-Lingual Evaluation Understudy*) – специална оценка, предложена от екип на IBM [*Papineni&al.,2002*], която измерва доколко машинният превод е близък на ниво 1, 2, 3 и 4 последователни думи (n-грами) до един или повече еталонни човешки превода. Пресмятанията се извършват за целия текст по следната формула:

$$BLEU = BP.\exp\left(\sum_{n=1}^{4}\frac{\log p_n}{n}\right)$$

където p_n е n-грамна точност спрямо еталоните, а $BP = \exp(\max(0,1-r/c))$ е наказание за прекалено къс превод (c и r са съответно дължините на машинния превод и на най-късия текст измежду еталоните).

Езиковият модел P(*e*)

Както споменахме по-горе, езиковият модел P(e) показва колко е вероятно да бъде казано дадено английско изречение e. Ако разгледаме e като последователност от думи $w_1w_2\ldots w_n$, можем да представим съответната вероятност като (4), което е еквивалентно на (5) съгласно правилото на веригата. Тъй като нямаме достатъчно данни, за да научим всички вероятности от (5), можем да направим приближението (6), което представлява Марковски модел от ред 2.

$$\begin{aligned} P(e) &= P(w_1w_2...w_n) && (4)\\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1w_2)P(w_4 \mid w_1w_2w_3)P(w_5 \mid w_1w_2w_3w_4)... && (5)\\ &\approx P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)P(w_4 \mid w_3)P(w_5 \mid w_4)... && (6)\end{aligned}$$

Марковският модел от ред n приближава вероятността на дадено изречение като произведение от условни вероятности на всяка дума при условие предходните $n-1$. Така например, при $n = 2$ вероятността на "*I eat an apple*" се пресмята така:

P("*I eat an apple*") = P(*I* | <S>) . P(*eat* | *I*) . P(*an* | *eat*) . P(*apple* | *an*) . P(</S> | *apple*)

В горната формула <S> и </S> са специални символи съответно за начало и за край на изречение и са необходими при моделиране на последователности с различни дължини.

Условната вероятност $P(w_i \mid w_{i-1})$ се пресмята най-общо като отношение на броя на срещанията $C(w_{i-1}w_i)$ на последователността от думи „$w_{i-1}\ w_i$" и на срещанията $C(w_{i-1})$ на думата w_{i-1} в голям обем текстове на английски език:

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i)}{\sum_{w_i} C(w_{i-1}w_i)} = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$$

На практика се използват по-сложни формули, които умеят да се справят с нулеви числители/знаменатели, както и с редки думи. Най-простото, но и най-неефективно решение е добавяне на някаква константа ε ($\varepsilon > 0$), например $\varepsilon = 1$, към всички $C(w_{i-1}w_i)$, преди използването им във формулата. Така получаваме следния вариант:

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i) + \varepsilon}{\sum_{w_i}[C(w_{i-1}w_i) + \varepsilon]} \approx \frac{C(w_{i-1}w_i) + \varepsilon}{C(w_{i-1}) + \varepsilon \mid V \mid}$$

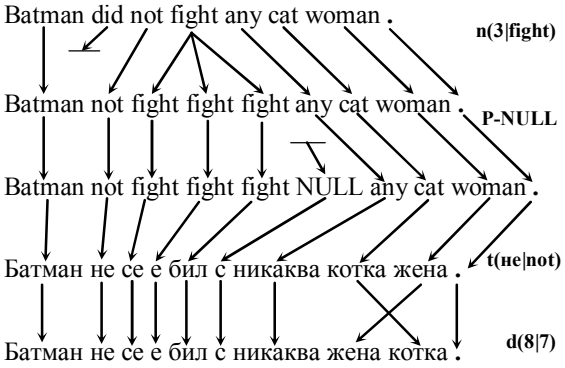
където $|V|$ е броят на различните думи в текста.

Обзор на най-популярните алтернативи, с подробна теоретична и практическа оценка, може да бъде намерен в [*Chen&Goodman,1999*].

Моделът на превода P(*b*|*e*)

Моделът на превода P(*b*|*e*) се учи от двуезичен паралелен корпус от български изречения и техния английски превод. Директното моделиране на P(*b*|*e*) в общия случай е невъзможно, тъй като е почти изключено да сме видели изреченията *e* и *b* в тренировъчния корпус, особено в случай на по-дълги изречения. Това налага представяне на P(*b*|*e*) с помощта на вероятности за превод на по-малки фрагменти.

Превод дума-по-дума



Фигура 4. Модел 3 на IBM.

Фигура 5. Подравняване на ниво думи.

Някои възможни решения се дават от класическите модели за статистически машинен превод, разработени от IBM през 1991 г. В статия от 1993 г. са публикувани 5 модела [*Brown&al.,1993*], по-късно станали известни като *модели на IBM* 1, 2, 3, 4 и 5. С цел простота на изложението по-долу ще изложим основната идея на модел 3 на IBM[1]. Това е генеративен модел, описващ процеса на трансформация на английско изречение в българско като поредица от четири стъпки, с всяка от които е асоциирана съответна вероятност. Процесът е илюстриран на Фигура 4, където е показано преобразуването на „*Batman did not fight any cat woman*.” в „*Батман не се е бил с никаква жена котка*.”. На първата стъпка се решава с колко думи ще се превежда всяка една английска дума: 0,1,2..., което се контролира от съответна вероятност $n(k|e_i)$, моделираща плодовитостта на съответната английска дума e_i. В нашия случай, спомагателният глагол *did* изчезва, глаголът *fight* се превежда с три български думи, а останалите английски думи ще се превеждат на български с по една дума. На втората стъпка на някои позиции се вмъква празната дума: това става с еднаква вероятност P-NULL, независеща от думите и позицията. В нашия

[1] През 2003 г. Ох и Ней предлагат модификация на Модел 5, която наричат Модел 6 [*Och&Ney, 2003*]. Модели 5 и 6 представляват по-скоро теоретичен интерес и не се използват в реални системи.

случай се вмъква празна дума, която по-късно ще съответства на предлога *c*, за който няма съответна дума в английското изречение. На третата стъпка всяка английска дума e_i се превежда с единствена съответна българска дума b_j съгласно вероятността за превод $t(b_j|e_i)$. Накрая, някои от българските думи се разместват съгласно вероятност $d(j|k)$: в нашия случай се разместват *жена* и *котка*. Четирите вероятности лесно могат да се научат, ако за всяка двойка изречения в двуезичния корпус имаме съответно подравняване на ниво дума като на Фигура 5. В общия случай обаче такова подравняване нямаме и се налага автоматичното му получаване с помощта на специален алгоритъм за максимизиране на очакването, даден в [*Brown&al.,1993*].

Таблици 2 и 3 показват примерни преводи, научени от модел 3 на IBM за английските думи *main* и *farmers* заедно със съответните вероятности за превод в проценти. В Таблица 2 се вижда, че *main* се превежда с форми на прилагателните *основен* и *главен*: в единствено/множествено число, в мъжки/женски/среден род, с/без членуване. Таблица 3 е още по-интересна и показва, че моделът е успял да научи, че *farmers* може да се преведе освен като *фермери*, и като *земеделски/селскостопански/частни стопани/производители*.

Български превод	Вероятност (в %)
основни	34.09
основните	22.73
основната	13.64
основно	6.82
основното	6.82
главната	4.55
важните	2.27
главен	2.27
главните	2.27
главният	2.27
основният	2.27

Таблица 2. Български преводи на английската дума main, извлечени с модел 3 на IBM.

Български превод	Вероятност (в %)
фермери	82.76
стопани	6.37
селски	2.67
селскостопанските	2.36
земеделските	1.88
земеделски	1.80
частни	1.13
производители	1.02

***Таблица 3. Български преводи на английската дума* farmers*, извлечени с модел 3 на IBM*.**

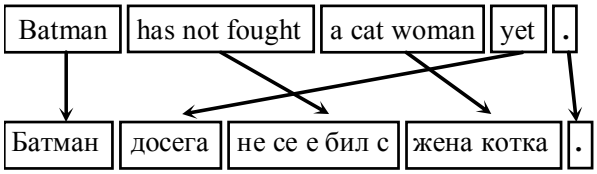
Важно ограничение на моделите на IBM е, че могат да учат само отношения от тип 1:M (едно към много), но не и M:1 (много към едно) или M:M (много към много), т.е. една английска дума може да поражда 0, 1 или повече български думи, но една българска дума може да се поражда само от една английска дума. Това означава, че ако на Фигура 4 вместо *any cat woman* имахме *the woman*, нямаше да можем да генерираме *жената* едновременно от *the* и от *woman* (както би било правилно), а щеше да трябва да приемем, че *woman* генерира *жената*, а *the* изчезва. Очевидно това създава трудности с определителния член при превод от английски на български с модел 3 на IBM, тъй като се губи връзката между членната форма в българското и в английското изречение.

Друг проблем е невъзможността за отчитане на контекста: напр. *interest rate* трябва да се преведе като *лихвен процент*, докато *interest in* – като *интерес към*. Неотчитането на контекста силно затруднява и съгласуването по род, число, падеж, членуване и др. и може да доведе до генериране на фрази като „*главен прокурорите*”, които невинаги могат да се коригират успешно от езиковия модел.

Превод с цели фрази

Преводът с цели фрази [*Koehn&al.,2003*] дава решение на повечето от проблемите, свързани с моделите на IBM. Това е генеративен модел, при който първо, английското изречение се разделя на “фрази”, след това, всяка фраза се превежда поотделно и накрая ня-

кои от фразите се разместват. С всяка от стъпките е асоциирана съответна вероятност, която се учи от паралелен корпус с български изречения и техните преводи на английски език. На Фигура 6 е даден конкретен пример.



Фигура 6. Превод с цели фрази..

Моделът води до съществено подобрение на качеството на машинния превод, тъй като умее да учи съответствия от тип M:M и дава възможност за по-добро отчитане на локалния контекст. Таблица 4 показва някои автоматично извлечени български фрази и техния английски превод. В първата част на таблицата се вижда, че моделът е научил, че думата *както* може да се превежда по различни начини в зависимост от контекста (например като *both*, *like*, *as*, *as well as*, *in line with*, *and* и др.) и е успял да идентифицира някои такива контексти.

Българска фраза	**Английска фраза**
както физическа , така и психическа	both physical and psychological
както целият регион	like the whole region
както те са определени	as defined
както и размера	as well as the size
както и предишните редовни доклади	in line with previous regular reports
както и по други	and in other
както и относно	and also about
както са	as were
както се предоставя на	as provided to
както следва :	as follows ,
главния	the base
главния	the chief
главния	the main
главния	the principal
главни прокурори	chief prosecutors
главни счетоводители	chief accountants
главни архитекти	chief architects
главни щабове	main staffs
главни улици	main streets
главни методисти	senior instructors
главно предизвикателство	major challenge

Таблица 4. Автоматично извлечени български фрази и техният английски превод.

Извлечените фрази не представляват непременно лингвистични единици и могат да включват препинателни знаци, което позволява на модела да отчете разлики в правилата за пунктуация между двата езика, например „*както следва :*” → „*as follows ,*”. Във втората част на таблицата виждаме някои възможни преводи на *главния*: както като отделна дума (напр. *the base*, *the chief*), така и в комбинация с други думи. Така например, *главни прокурори* (забележете, че в рамките на фразата има правилно съгласуване по род и число) се превежда като *chief prosecutors*, докато *главни методисти* е *senior instructors*, *главни улици* е *main streets*, а *главно предизвикателство* е *major challenge*. Накрая, вижда се, че за една българска фраза може да има множество възможни английски алтернативи за превод.

Както обяснихме по-горе, подходящият превод в конкретно изречение се избира като се взема предвид не само вероятността за превод на двойката фрази, но и вероятността за превод на цялото изречение според лингвистичния модел.

Забележете, че думите в Таблица 4 не са непременно в основната си форма – могат да бъдат в множествено число, членувани и т.н. Това е следствие от факта, че моделът не знае нищо за морфология, синтаксис, пунктуация и др. За него думата *главния* е толкова различна от *главният* и от *главен*, колкото и от *куче*, *зелени*, *играят*, *да, от* или , (запетая). Напълно възможно е моделът да знае фраза за правилен превод на *главни прокурори*, но не и за *главен прокурор*. По подобен начин, моделите на IBM може да могат да преведат *главния*, но не и *главен*, например защото никога не са го срещали в паралелния текст, върху който са обучавани. Това не е голям проблем при превод между английски и френски език (с които първоначално са експериментирали в IBM), но се оказва силно проблематично при превод между български и английски, тъй като българският е по-силно флективен, т.е. има повече различни форми на думите. Така например, за *главен* имаме *главен*, *главния*, *главният*, *главна*, *главната*, *главни*, *главните*, докато на английски за *main* има една-единствена форма, независима от род, число и членуване.

При превод с цели фрази най-често не се търси най-доброто произведение

$$\hat{e} = \arg\max_{e} \mathrm{P}(b \mid e)\mathrm{P}(e),$$

а се използва по-обща формула като

$$\hat{e} = \arg\max_{e} \mathrm{P}_{\text{фрази}}(b \mid e)^{\alpha_1}.\mathrm{P}_{\text{думи}}(b \mid e)^{\alpha_2}.\mathrm{P}_{\text{фрази}}(e \mid b)^{\alpha_3}.\mathrm{P}_{\text{думи}}(e \mid b)^{\alpha_4}.\mathrm{P}(e)^{\alpha_5}.\mathrm{length}(e)^{\alpha_6}... \qquad (7)$$

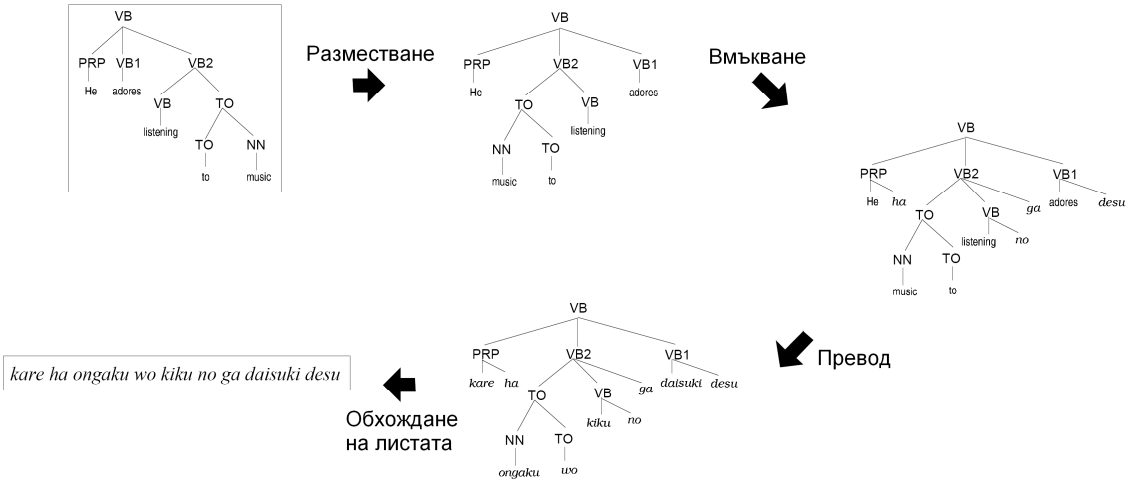
където $\mathrm{P}_{\text{фрази}}(x \mid y)$ и $\mathrm{P}_{\text{думи}}(x \mid y)$ са съответно вероятности за превод с цели фрази и дума-по-дума, *length(e)* е брой думи в английския превод (за да се избегне проблемът, че по-късите преводи по принцип имат по-голяма вероятност), а α_i са тегла, които се избират така, че да се постигне максимално *BLEU* върху някакви допълнителни паралелни изречения [*Och,2003*].

Превод с използване на синтактична информация

Съгласуването по род, число и падеж, както и някои особености на словореда създават сериозни проблеми на описаните по-горе модели, непознаващи понятия като съществително, глагол, подлог и др. Всъщност те не знаят дори какво е *дума*[2]: за тях няма разлика между истинска дума и препинателен знак. Затова следващата голяма цел на статистическия машинен превод е прякото моделиране на граматично знание. Задачата не е лесна, отчасти защото автоматичният синтактичен анализ е труден сам по себе си: най-добрите синтактични анализатори за английски език работят с 91% точност за качествен вестникарски текст, какъвто са обучени да анализират [*Charniak&Johnson,2005*], но качеството пада значително при други видове текст, например медицински. Въпреки това, най-добрите съвременни системи за синтактичен превод вече са приблизително изравнени със системите за превод с цели/йерархични фрази, като в някои редки случаи са дори по-добри.

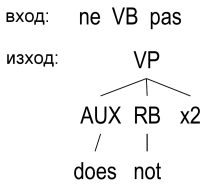
[2] Впрочем, при някои езици, напр. китайски и японски, определянето на границите между думите е тежка задача поради липсата на писмени разделители между думите и силната многозначност на различните комбинации от последователни йероглифи. Нужда от определяне на истинските граници на думите има и при някои европейски езици. Например, в италиански местоименията често се изписват залепени към края на глагола, напр. *compramelo = compra+me+lo* (‘*купи ми го*’), а в немски се сливат сложните съществителни, например *Stammzell* (т.е. ‘*стволова клетка*’; за сравнение в английски език се пишат разделени: *stem cell*).

Един теоретично изчистен ранен пример за синтактичен превод е моделът на Ямада и Найт [*Yamada&Knight,2001*], който описва процеса на трансформация на английско синтактично дърво в японско на три стъпки: (1) разместване на листа и поддървета, (2) вмъкване на допълнителни възли, и (3) превод на английските листа на японски език дума по дума. Накрая японското изречение се прочита от листата на дървото. Конкретен пример е показан на Фигура 7. Моделът е изключително подходящ за превод от английски на японски език, при което се налага вмъкване на множество допълнителни думи: това става на строго определени синтактични позиции, което прави синтактичния анализ абсолютно необходим.



Фигура 7. Превод от английски на японски език със синтактичния модел на Ямада & Найт.

Моделът на Ямаха и Найт започва със синтактично дърво на изходния език (английски), което преобразува в изречение на езика, към който се превежда (японски), т.е. моделът преобразува синтактично дърво в последователност от думи.

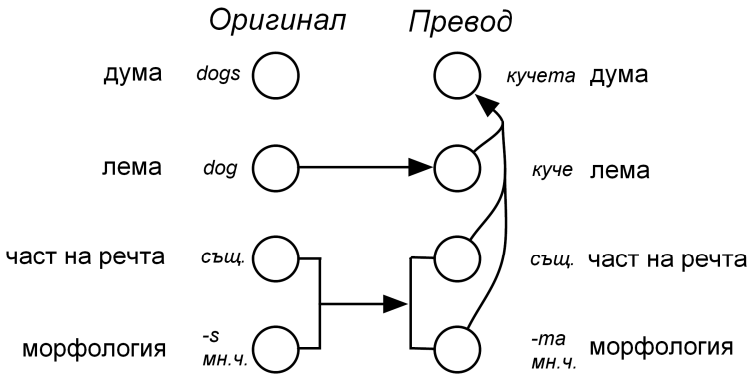


Фигура 8. Примерно правило за превод от френски на английски език с GHKM. *Правилото превежда стандартното френско отрицание с подходяща английска синтактична структура.*

Точно обратно работи моделът GHKM (по имената на авторите му Гали, Хопкинс, Найт и Марку) [*Galley&al,2004*], който по входна последователност от думи строи синтактични дървета за езика, към който се превежда. За целта GHKM използва правила като това на Фигура 8, с помощта на които се строят пълни синтактични английски дървета, които с много голяма вероятност съответстват на граматично правилни изречения.

Перспективна алтернатива на гореизложените синтактични подходи са факторните модели [*Koehn&Hoang,2007*], които позволяват просто моделиране на морфологични и лексикални характеристики на ниво отделна дума. Фигура 9 показва примерен превод на англ. *dogs* като *кучета*. Процесът включва три стъпки: анализ, транслиране и генериране. Първо, формата *dogs* се анализира като съществително *dog* в множествено число и съответно окончание *-s*. След това, лемата (основната форма), синтактичната и морфоло-

гичната информация се транслират поотделно в английските си еквиваленти. Накрая, те заедно генерират правилната българска форма *кучета*.



Фигура 9. Примерен превод на* dogs → кучета *с използване на факторен модел.

Основно предимство на факторните модели е, че позволяват отделен лингвистичен модел за всеки фактор, например на ниво морфология (за правилно съгласуване по род и число), на ниво част на речта (за граматично правилна последователност), на ниво лема (за семантично правилна последователност) и на ниво дума (за допълнително изглаждане). Тези лингвистични модели могат да се използват едновременно като отделни множители в обобщената формула (7), в резултат на което се получават граматично по-правилни преводи. Това обаче е свързано със значително общо забавяне на системата за машинен превод, което прави факторните модели приложими само при малки обеми тренировъчни данни.

Използваем ли е машинният превод?

Макар и далеч от качеството на професионалния човешки превод, днес машинният превод е вече използваем в ситуации, в които е достатъчно предаването на най-общия смисъл на текста, например при разглеждане на страница на чужд език в Интернет. Неслучайно функцията за автоматичен превод на *Google* е най-използваната сред предлаганите от компанията.

Автоматичният превод има какво да предложи и на професионалния преводач, например почти двукратно ускоряване на процеса на превод чрез подходящи подсказки в така наречените *преводачески памети* (у нас най-популярна е *Trados*). За някои хора-преводачи редактирането на компютърен превод се оказва още по-добра алтернатива, тъй като спестява нуждата от ръчно набиране на по-голямата част от текста. В Западна Европа и САЩ има компании, специализирали се в редактирането на компютърен превод, които предлагат висока скорост и качество на конкурентна цена. Накрая, но не на последно място, компании като *Xerox* от години превеждат цялата си техническа документация напълно автоматично и без никаква човешка намеса, благодарение на опростения език (*Caterpillar English*), който използват при съставянето на английския оригинал на документацията си.

Заключение

С разработването на статистическия подход през 1991 г. настъпи революция в областта на машинния превод, последвана от нова революция през 2003 г., когато беше предложен ефективен модел за превеждане с цели фрази. Оставаме в очакване на следващата голяма революция, която трябва да постави синтаксиса, морфологията, и граматичното знание изобщо, на полагащото им се централно място в процеса на машинен превод.

Литература

[*Brown&al.,1993*] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. *The mathematics of statistical machine translation: parameter estimation.* Computational Linguistics, 19(2), pp. 263-311, 1993.

[*Charniak&Johnson,2005*] E. Charniak, M. Johnson. *Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking.* In Proceedings of ACL. pp. 173-180, 2005.

[*Chen&Goodman,1999*] S. Chen, J. Goodman, *An empirical study of. smoothing techniques for language modeling*, Computer. Speech and Language, vol. 13, pp. 359-394, 1999.

[*Galley&al,2004*] M. Galley, M. Hopkins, K. Knight, and D. Marcu. *What's in a translation rule?* In Proceedings of HLT-NAACL'2004. pp. 273-280, 2004.

[*Hutchins,2003*] J. Hutchins. *ALPAC: the (in)famous report.* Readings in machine translation, ed. S.Nirenburg, H.Somers and Y.Wilks. pp. 131-135., The MIT Press, Cambridge, MA, 2003.

[*Koehn&al.,2003*] P. Koehn, F.-J. Och, and D. Marcu. *Statistical phrase-based translation.* In Proceedings of HLT/NAACL, Edmonton, Canada, 2003.

[*Koehn&Hoang,2007*] P. Koehn, H. Hoang. *Factored Translation Models*. Proceedings of EMNLP-CoNLL'2007, pp. 868-876, 2007.

[*Nagao,1984*] M. Nagao. *A framework of a mechanical translation between Japanese and English by analogy principle.* In: A.Elithorn and R.Banerji (eds.) Artificial and human intelligence (Amsterdam: North-Holland), pp. 173-180, 1984.

[*Och,2003*] F.-J. Och. *Minimum error rate training in statistical machine translation*. In Proceedings of ACL, pp. 160-167, 2003.

[*Papineni&al.,2002*] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. *Bleu: a Method for Automatic Evaluation of Machine Translation*. In Proceedings of ACL, pp. 311–318, 2002.

[*Weaver,1955*] W. Weaver. *Translation (1949).* In: Machine Translation of Languages, MIT Press, Cambridge, MA, 1955.

[*Yamada&Knight,2001*] K. Yamada, K. Knight. *A Syntax-based Statistical Translation Model*. In Proceedings of ACL, pp. 523-530, 2001.