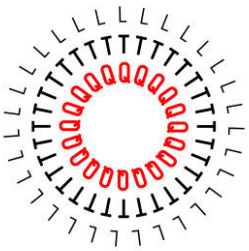
Европейски ден на езиците, 26 септември

Проектът QTLeap с хибридна технология за по-добри резултати в машинния превод



В последно време хиляди бежанци пристигат в Европа. Една от най-спешните задачи за тяхното приобщаване е те да научат езиците на страните, в които се озовават. Мащабът на тази задача следва да ни припомни, че предизвикателството на европейската интеграция все още не е истински преодоляно: езиковите бариери остават най-упоритите граници на Европа, като пречат на търговията, комуникацията и приобщаването.

От 2001 г. насам на 26.09. по инициатива на Съвета на Европа ежегодно се чества Европейският ден на езиците[1].

Ученето на езици не е лесна задача. Децата учат собствения си език или чужди езици, докато играят или слушат възрастните, но за възрастните ученето на граматически правила и нови думи е голямо усилие.

Компютърните програми като популярни онлайн системи за машинен превод учат езици подобно на децата – чрез това, което 'виждат' и 'чуват'. Те обаче използват така наречените методи за машинно учене върху големи количества двуезичен текст ("Big Data"), за да получат статистическите вероятности как определени думи и поредици от думи следва да се преведат от един език на друг. Така тези системи могат да предложат приблизителен превод, чието качество варира в зависимост от качеството и количеството на постъпващата информация.

Езикът, разбира се, е много повече от проста поредица думи. Той има граматически структури с елементи, каквито са сказуемите, подлозите и допълненията. Съществуват форми на машинен превод, базирани на лингвистични принципи, които използват подобен вид информация. Те, от своя страна, невинаги са подходящи за общи цели, но пък имат своите ниши на използване в различни конкретни области. Например, испанският вестник „Ла Вангардия" се превежда всяка вечер на каталунски с помощта на подобна система и с минимална човешка намеса за корекции, преди да бъде отпечатан.

В Интернет съществуват много ресурси (например вариантите на Уикипедия за машинно четене). Те съставляват така наречената Семантична мрежа и предоставят познания по семантика (значението на думите). Тези ресурси разкодират структурното знание за света. Например, те „знаят", че Париж е столица на Франция, но и че съществува град Париж в Тексас.

Институтът по информационни и комуникационни технологии (ИИКТ, БАН), който работи като български партньор по проекта на Европейския съюз QTLeap (qtleap.eu), изгражда хибридна технология за машинен превод, която обогатява

[1] http://edl.ecml.at/Home/tabid/1455/language/bg-BG/Default.aspx

статистическите методи с лингвистични познания и семантична информация от Интернет, с цел да произведе по-добри преводи.

„За да стане възможен добрият автоматичен превод, ни трябва много знание – не само лингвистично, но и за света" – казва доц. Кирил Симов, ръководител на българския екип. „Това означава, че са необходими много езикови и семантични ресурси. В момента използваме всички свободно достъпни източници на знание, като например Уикипедия, но за малките езици данните не са в такива мащаби, както за големите, затова се налага да се създават нови ресурси или да се допълват съществуващите ".

Както днес децата ни се чудят как преди двадесет години сме си писали домашните без Уикипедия, може би внуците ни ще се чудят как през 2015 г. всички хора, дошли в Европа в търсене на по-добър живот, са успявали да общуват преди създаването на *„преводача, който превежда всичко"*.

За повече информация и контакти: qtleap.eu.