Изкуствен интелект—Зимен семестър 2001/2002. Преподавател: Иван А. Держански.

Лекция 11: Извличане на информация

извличане на информация (information extraction): процес, при който от непознат текст (или текстове) се получава еднозначна информация, представляваща интерес според зададени критерии. За целта обикновено се използуват

- крайни автомати и други методи от обработката на естествен език,
- системи от знание за реалния свят,
- форми на машинно самообучение.

Има връзка и сходство с

- **автоматичното резюмиране (text summarisation)**, само че критериите за подбор на информацията са зададени от потребителя във формата на **шаблони (templates)**;
- **търсенето на информация (information retrieval)**, само че резултатите се привеждат в предварително определен фиксиран формат. След това те може да се предоставят непосредствено на потребителя, да се запазят в база данни или електронна таблица или да служат за резюмиране или за индексиране и класифициране на документи за нуждите на търсенето на информация. Този етап е сходен с локализацията на програмни продукти;
- **отговарянето на въпроси (question answering)**, което обаче изисква готова база данни;
- **машинния превод (machine translation)**;
- неограниченото **разбиране на текст (text understanding)**, което още не е постигнато.

Ефективността на извличането на информация като цяло зависи от

- езика (повечето изследвания са върху английски, японски, испански и китайски);
- стила, жанра и предметната област на текста;
- вида сценарий (описания на състояния или на минали събития, инструкции), от който се интересува потребителят.

Широкомащабна изследователска област от края на '80-те години. Напредъкът се изразява в

- работа с по-разнообразни и по-сложни текстове,
- по-точно определени задачи (вж. следващия раздел),
- по-добри критерии за оценяване на точността на процеса,
- повишаване на модулността на системите и приспособимостта им към други приложения.

1 Задачи на извличането на информация

1.1 Разпознаване на именувани индивиди

... или за какви индивиди (от различни типове: хора, места, организации, транспортни средства, суми пари, дати и т. н.) става дума в текста. За всеки индивид се създава запис, съдържащ служебно име (идентификатор) и означение за типа.

Точността (засега за английски език) достига около 95%, което значи, че е съпоставима с човешката; поради това технологията има много приложения. Зависи донякъде от предметната област.

1.2 Разрешаване на кореферентности

... или кои изрази (описания, местоимения) за кои индивиди се отнасят. Не е от непосредствен интерес за потребителя, а само съпътства другите задачи. Позволява да се събере разхвърляната из текста описателната информация за един индивид.

Точността не е много по-висока от 50%, макар че е различна за откриването на кореферентност на собствени имена и за значително по-сложното разрешаване на анафорични (местоименни) връзки. Зависи от предметната област.

1.3 Изграждане на шаблонови елементи

... или какви атрибути имат индивидите. Открива и свързва с тях описателна информация (вкл. имена в текста). Точността на най-добрите системи е около 80% — доста под човешката. Зависи донякъде от предметната област.

1.4 Изграждане на шаблонови отношения

... или какви отношения (вкл. състояния и събития) от интерес съществуват между индивидите — шаблонови елементи. Централна част от почти всяка задача за извличане на информация.

1.5 Създаване на шаблони за сценарии

... или в какви събития участват индивидите. Шаблоните за сценарии са типичният резултат от работата на системите за извличане на информация. Те свързват шаблоновите елементите в описания на отношения и събития, към които може да се добави произволна описателна информация (достигнат етап, източник на сведенията и т. н.).

Точността на най-добрите системи е около 60%, но и човешката не е много над 80%. (Може да се увеличи, но не с много, за сметка на ефективността.) Зависи от предметната област и от сценариите.

2 Многоезиково извличане на информация

Засега относително малко изучен въпрос. Между английския или испанския език от една страна, японския от друга и китайския от трета има много малко сходства, но те са съществени (думите нямат много различни форми, словоредът е повече или по-малко фиксиран). Експериментите с езици, по-сложни в граматично отношение, едва сега започват.