

Лекция 11:
Машинно самообучение (продължение)

САМООБУЧЕНИЕ ЧРЕЗ ОБЯСНЕНИЕ

Идея. При самообучението чрез примери съществено е честото използване на голямо множество от обучаващи примери (положителни и отрицателни). Тук целта е да се минимизира множеството от обучаващи примери, като по възможност се използва един (положителен) обучаващ пример и на основата на допълнителни знания за предметната област се конструира обяснение защо той е пример за разглежданото понятие. Чрез обобщаване на обяснението се получава описанието на разглежданото понятие.

При МС чрез обяснение съответната обучаваща програма използва следните данни:

- *Целево понятие* (описание на целевото понятие в термините на предметната област, но без да е удовлетворен т. нар. критерий за операционалност);
- *Критерий за операционалност (конструктивност)* – описание или списък на понятията, в чиито термини трябва да бъде формирано търсеното описание на целевото понятие;
- *Обучаващ пример* – описание на положителен обучаващ пример за целевото понятие;
- *Теория за предметната област* – множество от правила, които описват връзките между обектите и понятията в предметната област (база от знания за предметната област).

При това, обучаващият пример и теорията за предметната област би следвало да бъдат описани така, че да удовлетворяват критерия за операционалност.

На основата на тези данни обучаващата програма конструира обяснение (доказателство) за това, че обучаващият пример е описание на обект - представител на целевото понятие. Това обяснение (доказателство) се обобщава и така се получава търсеното описание на целевото понятие, което удовлетворява критерия за операционалност.

Най-съществено за МС чрез обяснение е използването на знания за предметната област. Това е характерна черта на този метод, която го отличава от всички останали. При това положение естествено възниква въпросът, за какво е необходимо използването на обучаващ пример за разглежданото понятие, след като по принцип знанията за предметната област са достатъчни за извършването на класификация на обектите в нея. Отговорът е, че по този начин се постига по-голяма ефективност и конструктивност. Примерът насочва и ограничава знанията за предметната област, които са приложими в конкретния случай. Иначе в процеса на извод ще се използва цялата база от знания за областта.

РЕШАВАНЕ НА ЗАДАЧИ И САМООБУЧЕНИЕ ЧРЕЗ АНАЛОГИЯ

Идея. Това е метод за решаване на задачи и МС, съответен на човешкия подход за решаване на нови задачи с помощта на аналогии с неща, за които е известно как се

правят. Този процес е значително по-сложен от конструирането и запазването на макрооператори, тъй като старата задача може да бъде съвсем различна на външен вид от новата, която се опитваме да решим. Основната трудност тук е в определянето на това, кои неща са подобни и кои не са. За създаването и използването на аналогии се изисква формулиране на подходяща *метрика*. Най-често използваният метод за решаване на задачи по аналогия е *трансформационната аналогия*.

Същност на трансформационната аналогия. Ако трябва да се реши задача в дадена област, търси се вече решена задача, чиято формулировка е подобна (аналогична) на формулировката на новата задача, след което се трансформира нейното решение, като при необходимост се правят подходящи субституции и евентуално доуточняване на полученото решение (след допълнително търсене в т. нар. пространство на решенията).

ИЗВЛИЧАНЕ НА ИНФОРМАЦИЯ

Същност на извличането на информация (Information Extraction, IE): процес, при който от непознат текст (или текстове) се получава еднозначна информация, представляваща интерес за потребителя според зададени критерии.

За целите на извличането на информация обикновено се използват:

- крайни автомати и други инструменти и методи от обработката на естествен език;
- системи от знания за реалния свят;
- форми на машинно самообучение.

Извличането на информация има връзка и сходство с:

- **автоматичното резюмиране (text summarization)**, само че критериите за подбор на информацията са зададени от потребителя под формата на *шаблони (templates)*;
- **търсенето на информация (information retrieval)**, само че резултатите от търсенето се привеждат в предварително определен фиксиран формат. След това те може да се предоставят непосредствено на потребителя, да се запазят в база от данни или електронна таблица или да служат за резюмиране или за индексирание и класифициране на документи за нуждите на търсенето на информация;
- **машинния превод (machine translation)**;
- **неограниченото разбиране на текст (text understanding)**, което едва ли скоро ще бъде постигнато.

Ефективността на извличането на информация като цяло зависи от:

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

В литературата обикновено се разглеждат пет типа **задачи на извличането на информация**: *разпознаване на именувани индивиди* (Named Entity recognition, NE), *разрешаване на кореферентности* (Coreference resolution, CO), *изграждане на шаблонни елементи* (Template Element construction, TE), *изграждане на шаблонни отношения* (Template Relation construction, TR), *създаване на шаблони за сценарии* (Scenario Template production, ST).

Разпознаване на именувани индивиди

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

Точността зависи донякъде от предметната област, но като цяло (засега за английски език) достига около 95%, което означава, че е сравнима с човешката. Затова разглежданата технология има много приложения.

Разрешаване на кореферентности

Определяне на това, 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 описателна информация за един индивид.

Точността не е по-висока от 50%, макар че е различна например за откриването на кореферентност на собствени имена и разрешаването на местоименни връзки (последното е значително по-сложно). Зависи от предметната област.

Изграждане на шаблонни елементи

Определяне на това, какви атрибути имат индивидите. Открива се и се свързва с тях описателната информация в текста. Точността на най-добрите системи е около 80% - доста под човешката. Зависи донякъде от предметната област.

Изграждане на шаблонни отношения

Определяне на това, какви отношения (вкл. състояния или събития) от интерес за потребителя съществуват между индивидите – шаблонни елементи. Това е централната част от почти всяка задача за извличане на информация.

Създаване на шаблони за сценарии

Определяне на това, в какви събития участват индивидите. Шаблоните за сценарии са типичният резултат от работата на системите за извличане на информация. Те свързват шаблонните елементи 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%. Зависи от предметната област и от сценариите.