Word-Sense Disambiguation (WSD)

Учихме за POS-tagging - автоматично разпознаване на думите като части на речта в непознат текст - без пълен анализ, но със специални процедури, които са обучени да предлагат най-вероятната конфигурация от линейни последователности на категориите. Това се прави от програми, наречени POS-taggers. Пример за българския език: знаем, че низът БЕЛИ може да е форма на (1) глагол, (2) прилагателно и (3) съществително. Идва непознат и неанализиран текст. Успешен POS–tagging предполага системата да разпознае кое от трите е низът БЕЛИ в конкретните му появи в контекста. Но не можете да очаквате, че с POS-tagging ще разпознаете например дали БЕЛИ е заповедна форма или минало време (ти бел'и) на глагола *беля*, когато се случи БЕЛИ да се разпознае като глагол. Такива детайли не влизат във фокуса на задачата за POS-tagging, понеже обучението не се провежда спрямо толкова детайлно описани категории. Казахме, че броят на категориите трябва да бъде сравнително скромен, тъй като при много маркери задачата не може да се реши в реално време.

Сега минаваме към друга най-основна процедура при обработката на непознат текст: разпознаване на смисъла на думите без пълен анализ. Например, знаем от някакъв семантичен ресурс (речник, интегриран в системата), че КОСА като съществително има две значения. Задачата е да разпознаем кое от двете се среща в текста при конкретните употреби на низа КОСА като съществително. Пълен анализ би ни помогнал - понеже семантичната интерпретация би филтрирала правилното значение - но това на практика е невъзможно. Използват се статистически техники; няма най-добра или универсална; задачата е (все още) на етапа на експерименталната оценка на поведението на различни алгоритми с различни входни данни.

1. Постановка на задачата

Задачата за WSD е по същество класификационна: даден е набор от значения (т.е. семантични маркери в някакъв речник) за определени думи; можем ли да разберем кой е най-подходящият маркер за всяка индивидуална употреба на думата в текста? Наборът значения - т. нар. семантични етикети - е от съществено значение за естеството и сложността на WSD-задачата. Примерни ресурси от такива семантични етикети за английското съществително **sentence** са дадени по-долу в Таблица 1.

Речник COBUILD	Превод на испански	Предметна област	Употреба в контекста
Noun-2	sentencia	право	... for a maximal *sentence* for a young ...
Noun-2	sentencia	право	... of the minimum *sentence* of seven years ...
Noun-2	sentencia	право	... were under the *sentence* of death ...
Noun-2	sentencia	право	... criticize a *sentence* handed down ...
Noun-1	frase	лингвистика	... in the next *sentence*, they say ...

Noun-1	frase	лингвистика	... read the second *sentence* because ...
Noun-1	frase	лингвистика	... as the next *sentence* is an important ...
Noun-1	frase	лингвистика	... is the second *sentence* which I ...
Noun-1	frase	лингвистика	... listen to this *sentence* uttered by a ...

В първите 3 стълба виждаме 3 отделни семантични ресурса с три вида етикети. В първия стълб е "номерът на значението" според известния речник COBUILD за англ. език (издаден 1987). Тоест имаме налице електронен речник, в който някой е записал на ръка по колко значения има всяка английска дума като някаква част на речта (така е организиран почти всеки добър речник, ще го забележите ако разгледате някой едноезичен български тълковен речник - има няколко на хартия, но за жалост при доста речници тази информация е неформатирана в текста и не се отделя лесно и нееднозначно). Вторият стълб е ресурс тип "превод на друг език". Често преводите дават представа за значенията, понеже при тях проличава отделното значение. Така задачата за разпознаване на семантичната многозначност при машинния превод например прилича на задачата "как да се преведе дадена дума". В третия стълб стои етикет на значението според предметната област. Често такива етикети са налични в речници-енциклопедии, които не номерират значенията, а ги изброяват едно след друго с указател за предметната област. Макар тези семантични ресурси да са събрани заедно в таблица 1, на практика най-често е достъпен един от тях и се работи само с него. Оказва се, че трите семантични ресурса в стълбове 1-3 имат почти еднаква "разрешителна сила" за разпознаване на употребите от стълб 4 (при използваните днес методи). Да повторим, че WSD означава да се закачи tag на низа в стълб 4, т.е. един от етикетите в стълбове 1,2 или 3 се присвоява на употребата на *sentence* в стълб 4 като най-вероятен. Най-общо това става чрез наблюдения на непосредствено околните думи (т.е. думите в стълб 4), които дават "най-тежка" мотивация за избор на значението. По-далечните окръжаващи думи (в същия параграф, същия документ) дават по-слаби доказателства.

Да изброим по-подробно видовете семантични ресурси (които са важни като начална информация в системата):

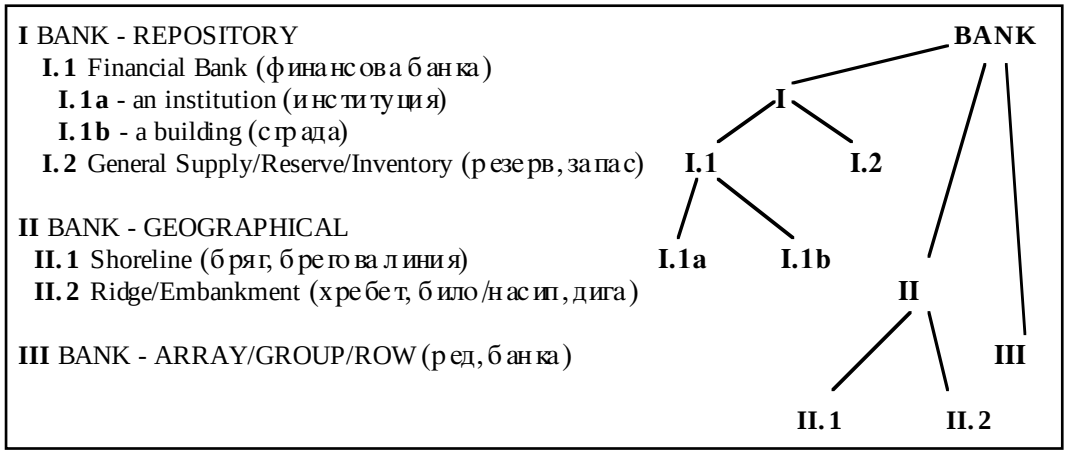
- речници от типа на COBUILD, пример стълб 1 на таблица 1. Използват се широко при WSD. Хубавото е, че в тези речници има примерни изречения за *типични* употреби и така системата трупа и учи характерни думи от непосредствения контекст;
- Концептуални йерархии от типа на WordNet (говорили сме за нея). Такъв ресурс позволява да се ползва наследяване в класа (class-based inheritance) и ограничения при филтриране на съчетаемостта (selectional restrictions);
- Кодове на предметната област, асоциирани към смисъла. Така е построен например *LDOCE (Longman Dictionary Of Contemporary English, 1978*, сега го има on-line). Пример за това се дава в стълб 3 на табл. 1.;

- Многоезикови преводни съответствия и полисемии (многозначности). Колкото повече езици ползвате, от различни семейства, толкова повече смислови многозначности можете да изразите (както в стълб 2 на табл. 1.). Така се подхожда при машинния превод, но не само там. Ако наблюдавате паралелни корпуси от вече преведени текстове на два езика, можете да мислите за автоматично извличане на данни подобни на тези в стълб 2 на табл.1. Тоест, можете да проверите дали дадена дума е семантично многозначна (в определена предметна област), понеже се превежда по различни начини. Възможността да се автоматизира събирането на данните е голямо предимство; демо за автоматичен екстрактор се прави на последната лекция.

- специализирани ресурси ad hoc. Строят се ръчно в тесни предметни области за отделни приложения. Учихме за MUC, представете си, че строите ръчно семантичен ресурс с цел да подпомогнете разпознаването на смисъла при думи, които искате да сложите в "шаблона" на системата.

Грануларност на отделяните значения

Отговорът на въпроса "*колко трябва да раздробим значенията, които искаме да ги разпознаем*" често зависи от приложението. Например, ако правим система за английско->френски машинен превод и двете значения на bank **I.1a** и **I.1b** (вж. фиг. 1) се превеждат с една и съща френска дума - то тогава защо да полагаме усилия системата да ги различава? Но очевидно това не е случая с превода "sentence" -> "изречение/присъда" от таблица 1; там неразпознаване на значенията води до грешен превод. Но по принцип: колкото са значенията в семантичния ресурс, толкова можем да разпознаваме (понеже няма как да научим системата на повече значения).



Фиг. 1. Различни значения на думата BANK на английски. Забележете, че само I и III се покриват със значенията на БАНКА на български

Важен въпрос е как мерим грешката при некоректно разпознаване на смисъла в текста спрямо базиса, зададен в семантичния ресурс. Изчислява

се т.нар. “наказателна матрица” (**penalty matrix**), в която се внасят “наказателни точки” за грешно присвоен смисъл в текста, при предварително зададена функционална семантична близост между всеки две значения, които подлежат на разпознаване. Таблица 2 е пример за *функционалната семантична близост* между значенията на bank, показани на фиг. 1. Тя е формирана на принципа:

	I.1a	I.1b	I.2	II.1	II.2	III
I.1a	0	1	2	4	4	4
I.1b	1	0	2	4	4	4
I.2	2	2	0	4	4	4
II.1	4	4	4	0	1	4
II.2	4	4	4	1	0	4
III	4	4	4	4	4	0

Табл. 2. Примерна семантичната близост между двойките значения на bank

- на разстояние 1 са наследници на второ (най-ниско) ниво вътре в главно значение (тоест това са значения с близост 1, *много близки*);
- на разстояние 2 са наследници на първо ниво на главно значение;
- на разстояние 4 са елементите на главните три значения.

При избора на тези числа-“семантични дистанции” се използват психо-лингвистични експерименти, които са насочени към изучаване на човешките грешки при определяне на семантичната дистанция.

След като имаме зададена табл. 2, можем да въведем по-чувствителни мерки за сериозността на грешката, например:

$$\text{точност на разпознаване на смисъла} = \frac{1}{N} \sum_{i=1}^{N} d(c_i, a_i),$$

където N е броят на направените тестове, c_i е коректният смисъл, a_i е присвоеният смисъл, и $d(c_i, a_i)$ е разстоянието между двете значения (коректно и присвоено) според таблица 2. Може и по-фино, ако се отчита думата, наличието на контекста и условната вероятност:

$$\text{точност на разпознаване} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{L} d(c_i, a_i) \times P(s_j \mid w_i, \text{context}_i),$$

където s_j , j от 1 до L, са различните значения на думата w_i.

2. Приложения на WSD

Поне четири ни идват веднага наум:

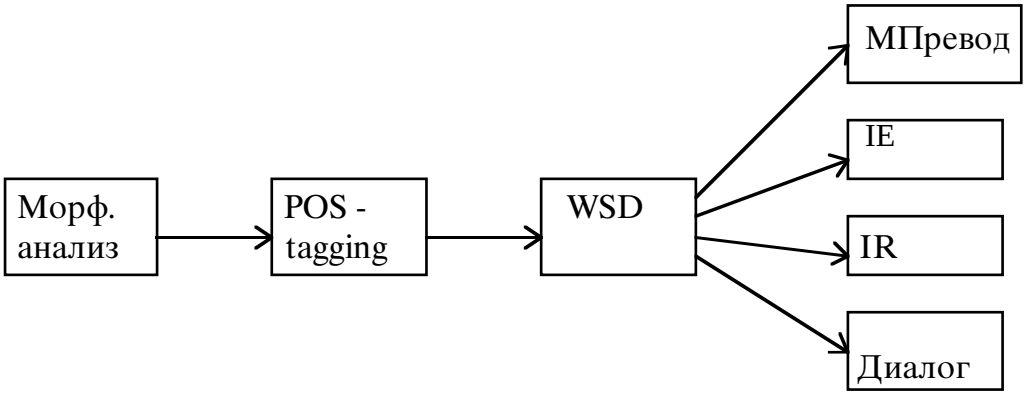
Машинен превод. Би било от голяма полза за една система да разпознае в кой смисъл е употребена някоя дума във входния език, за да знае как да я преведе на изходния език. Неразпознаване на смисъла е една честа грешка при днешните системи за МП.

Information Retrieval (IR). По идея, би следвало да очакваме, че разпознаването на смисъла на думите в текста би ни помогнало да мерим по-добре близостта между документите. Например, *tank* значи на английски както военно съоръжение, така и контейнер. Разликата е твърде голяма и много хора се надяват, че различаването на двата смисъла би подобрило съществено поведението на алгоритмите за различаване на двата вида документи. На практика обаче, поне засега, ползата от WSD при IR не се доказва експериментално. IR-системите с WSD имат съвсем малко по-добри резултати (от порядъка на 1%). Това се обяснява с факта, че WSD и IR експлоатират едни и същи лингвистични факти, например IR би свела военен документ за танкове до {Panzer, infantry, tank} - докато WSD би разпознала tank като военен термин, понеже се среща в контекста на Panzer и infantry. Така че WSD прилага *като ехо* контекстуалната информация, която вече е употребена в IR за присвояване релевантност или близост на документа към някаква предметна област. По тази причина комбинацията на двете техники е със съмнителен успех (засега).

Message Understanding (MUC), или *Information Extraction*. За този вид системи е от голямо значение да разпознаят значенията на думите, които сигнализират шаблони, за да започнат анализ на изречението. Например - ако системата различава *drug* в значенията му *лекарство* и *наркотик* - би могла да подходи избирателно към текста. Друг пример е необходимостта да се разпознаят имената: дали едно име е на лице, институция, географски обект и т.н. Към този проблем също може да се подходи с WSD.

Човеко-машинен диалог. Когато хората говорят на компютъра в диалогова система, също е важно да се разбере колкото се може по-бързо и “евтино” кои значения са употребени.

Картинката става както следва на фиг. 2:



Обърнете внимание, че ако искате да прилагате WSD за български или друг флективен език, първо трябва да се преборите с морфологичния анализ (т.е. да сведете всяка дума към основната й форма, което се нарича *лематизация*). Това е значително препятствие, което съществено затруднява обработката на флективни езици.

3. Ранни подходи към задачата за WSD

Споменаваме първия пример като христоматиен, от зората на машинния превод, когато подходът е бил rule-based и не се е прилагала сериозна статистика с тогавашните компютри. Коментарът в от 1960 година, когато е забелязано колко е безнадеждно да се различи автоматично значението *кошара* (а не *писалка)* от триграмата *in the pen* в изречението *the box is in the pen* (понеже и за двете значения се казва in). Едва през 90-те години се появяват машини, с които става възможно да се покаже, че от **вероятностна** гледна точка *in* и *pen* са много добри указатели за значението *кошара,* особено ако наоколо не е спомената дума от рода на *мастило*.

Поради затишието в обработката на естествен език през 60-те и 70-те години, дължащо се на финансови причини, ОЕЕ следва плътно ИИ за около 10-15 години. Тогава се появява парадигмата на "NL Understanding", каквато я видяхме на лекции 1-8 (т.нар. AI-NLP). Разпознаването на смисъла се извършва с умозаключения в процеса на семантичната и контекстуалната интерпретация.

След 1985 година започват да си пробиват път т.нар. data-driven техники, където основният инструмент е статистиката. От 90-години това става доминиращ подход и в момента все още е така.

4. Supervised-подходи към задачата за WSD

Както знаете, има два главни подхода при машинното самообучение: с учител (supervised learning) и без учител (unsupervised learning). При WSD е трудно да се разграничат двете - поради вариантите на прилагане на основните техники, но все пак главната разлика е както следва:

Supervised WSD (s–WSD) най-често си извлича правилата за класификация и/или статистическия модел директно от маркирани учебни примери на многозначни думи в контекст. Обикновено са необходими стотици примери за научаване на адекватната класификация и липсата на маркирани учебни корпуси е основен проблем за s-WSD.

За разлика от него, unsupervised WSD (us-WSD) не изисква предварително маркиран учебен корпус; в най-чистия си вид тези подходи извличат

информация за различните значения от суров текст. Някои подходи си "помагат" с използване на външни ресурси от типа на WordNet, поради което ги наричат minimally supervised.

А. Учебни данни за s-WSD

Споменаваме ги, за да добиете представа за размера на наличните ресурси. Има само няколко колекции анотирани данни, и то само за английски език.

Най-голямата е SEMCOR semantic concordance (1994->), където един едномилионен корпус прогресивно се маркира със значенията на думите в WordNet за всяка дума. SEMCOR покрива балансирано широк кръг думи, но има сравнително малко употреби за многозначните думи. Малък, но често използван е line-корпусът (1993), където 2094 примера на думата line са маркирани с 6-те основни значения. Трета налична колекция е корпус за 12 многозначни думи, с 300-11000 значения за всяка дума. Четвъртата колекция - маркиран корпус за 190-те най-чести многозначни думи на английски, с по 1000 примера на дума. Всички тези данни са публично достъпни от Lingustic Data Consortium (ldc@unagi.cis.upenn.edu).

Учебните данни се наблюдават като таблица. Табл. 3 илюстрира появата на маркираната дума plant с два маркера MANUFACT и LIVING в контекст от 3 околни думи (обикновено се работи с 50 околни думи).

Маркер на значението	Употреба на многозн. дума в контекст
MANUFACT	... from the Toshiba *plant* located in ....
MANUFACT	... union threatened *plant* closures ...
MANUFACT	... chloride monomer *plant*, which is ...
LIVING	... with animal and *plant* tissues can be ...
LIVING	... Golgi apparatus of *plant* and animal cell ...
LIVING	... the molecules in *plant* tissue from the ...

Табл. 3. Примери на маркирани многозначни думи в контекст

Наблюдават се околните словоформи, техните леми (основни думи), POS-таговете им, относителна позиция и синтактични функции (ако можете да пуснете парсер). Пример за представяне на "околните" данни има в табл. 4.

	Дума-2			Дума-1			Дума 0	Дума +1
Маркер	Клас	Дума	POS	Клас	Лема	POS	Дума	Дума
MAN...		the	DET	CORP	Toshiba	NP	Plant	Located
MAN...	BUSN	union	NN		threaten	VBD	Plant	Closures

MAN...	CHEM	chloride	NN	CHEM	monomer	NN	Plant	,
LIVING	ZOOL	animal	NN		and	CON	Plant	Tissues
LIVING		apparatus	NN		of	PREP	Plant	And
LIVING	CHEM	molecules	NNS		in	PREP	Plant	Tissue

Табл. 4. Данни за контекста на plant, запълнено е само известното

След получаване на табл. 4 може да започне наблюдение на честотите на разпределение на семнтичните маркери в зависимост от околните характеристики. Пример има в табл. 5: *f(M)* е честотата на появата на plant като MANUFACT, а *f(L)* - честотата като LIVING. Тази "сурова" статистика е основата за всички по-нататъшни класификационни решения.

Наблюдаван атрибут	Шаблон за наблюдение на атрибута	f(M)	f(L)	В повечето случаи е
word+1	*plant growth*	0	244	LIVING
word+1	*plant height*	0	183	LIVING
lemma+1	*plant size/N*	7	32	LIVING
lemma+1	*plant closure/N*	27	0	MANUFACT
word-1	*assembly plant*	161	0	MANUFACT
word-1	*nuclear plant*	144	0	MANUFACT
word-1	*pesticide plant*	9	0	MANUFACT
word-1	*tropical plant*	0	6	LIVING
POS+1	*plant* <NOUN>	561	2491	LIVING
POS+1	<NOUN> *plant*	896	419	MANUFACT
word±k	*car* в околните ±k думи	86	0	MANUFACT
word±k	*union* в околните ±k думи	87	0	MANUFACT
word±k	*job* в околните ±k думи	47	0	MANUFACT
word±k	*pesticide* в околните ±k думи	9	6	MANUFACT
word±k	*open* в околните ±k думи	20	21	LIVING
word±k	*flower* в околните ±k думи	0	42	LIVING
Verb/Obj	*close* /V, Obj = *plant*	45	0	MANUFACT
Verb/Obj	*open* /V, Obj = *plant*	10	0	MANUFACT
Verb/Obj	*water* /V, Obj = *plant*	0	7	LIVING

Табл. 5. Разпределение на честоти при наблюдение на *plant* в учебен текст

Забележете, че синтактичната позиция може да бъде от решаващо значение. Заводите се отварят и затварят, растенията се поливат. Всяка самообучаваща се програма би разгледала тези атрибути като 100% мотивация за вземане на решение как се категоризира plant. По-добрите системи за WSD наблюдават синтактичните позиции и контекста

структурирано, а не като неподреден набор от думи. НО: няма много такива усъвършенствани системи - от една страна, и от друга: поради липсата на учебни корпуси е затрудено сравняването на отделните системи (които наблюдават различни данни от типовете, показани на Табл. 1).

Най-общо поведение на s-WSD

Прилаганите s-WSD методи показват следното поведение:

- ***decision trees*** - около 70-75% точност при разпознаване на коректното значение, при 37% вероятност за случаен избор; друга система осигурява 22% подобряване на преводите в системата на IBM за френско-английски машинен превод (от 37 приемливи превода до 45 приемливи при 100 теста).

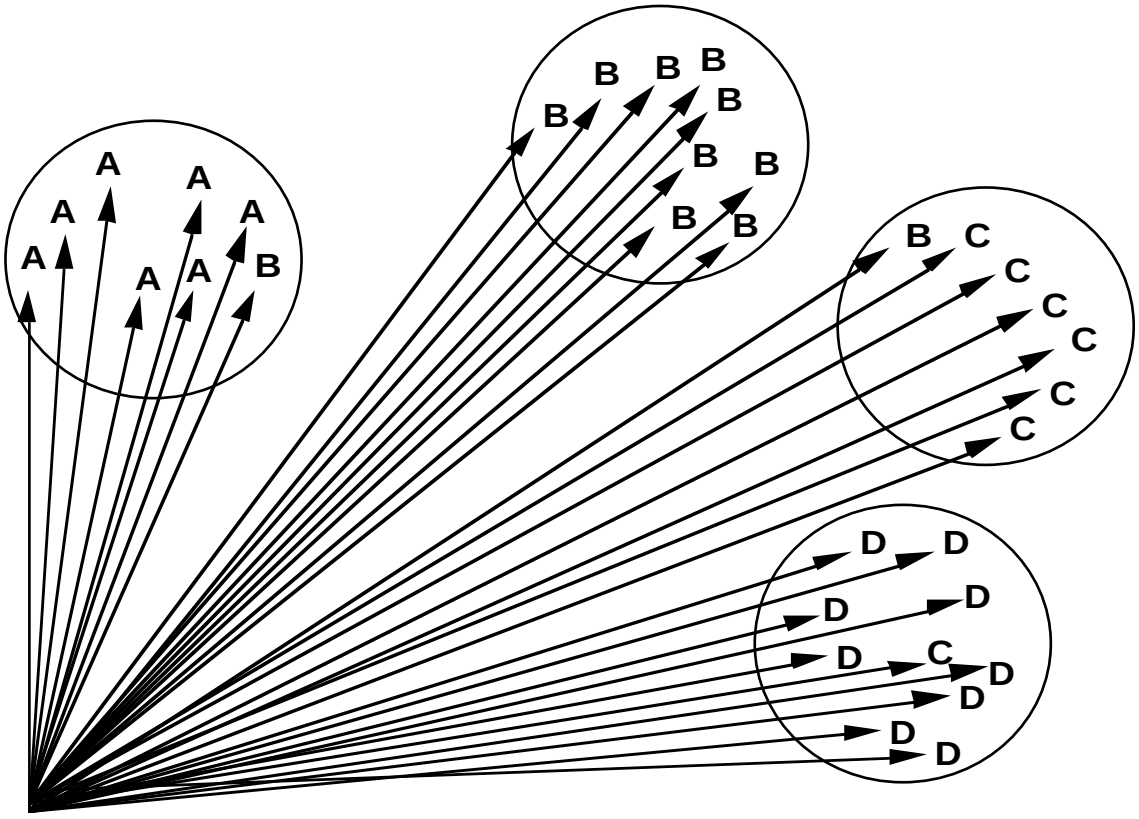
- Бейсов класификатор (**Bayesian Classifier**) - както е известно, това е един от най-старите алгоритми за машинно самообучение. За WSD е прилаган през 1992 за многозначни думи с по едно до две значения, над френско-английки паралелен "подравнен" корпус, с наблюдение на 50 думи отляво в контекста и 50 думи отдясно. За 6 многозначни думи (duty, drug, land, language, position, и sentence) със 17138 учебни примера е достигната средна точност 90%. Друга система - която наблюдава 12 многозначни думи, с друг специфичен набор от учебни атрубути - достига около 84% точност.

Други прилагани s-WSD методи са decision lists, content vector models, и т.н.

5. Unsupervised-подходи към задачата за WSD

Както казахме, има само няколко колекции учебни данни, и то за английски (или двуезични - но пак включващи английски). Поради това методите за WSD без учител имат шанса да се прилагат и за други езици.

Клъстери на значенията според векторно пространство: В текста за Information Retrieval се показа как документите се свеждат до вектори от думи. По същия начин "околните думи" в контекста на многозначните думи могат да се представят като вектори и да се формира пространство от значенията на думите в техния контекст. Близките вектори съответстват на "подобни" значения. Представете си една многозначна дума с четири значения A, B, C, D. Те не са известни по начало, но се забелязват от 4-те клъстера формирани от картинки подобни на показаната на фиг. 3. Както е споменато в текста за IR, получава се много голямо пространство и се прибягва до редуцирането му с използването на т.нар. singlular value decomposition. Важен проблем при този подход е, че без маркирани контекстни вектори няма как да се направи съответствие между намерените клъстери и семантичен ресурс от типа на WordNet.



Фиг. 3. Индуцирани клъстери от “близки” вектори, които съответстват на отделни значения на наблюдаваната дума

Прилагат се и други WSD алгоритми без учител: iterative bootstrapping, topic-driven class models, и hierarchical class models. Точността е 70 - 85%, но това са прототипи работещи със значенията на няколко думи.

/Забележка: bootstrap е метод за оценка на параметри на разпределение. Състои се в генериране на данни с разпределение определено от данните (емпиричното разпределение). С други думи от данните се вадят части, гледа се какво се получава като оценка и така много пъти, за да се получи представа за разпределението на оценката. За повече сведения доц. Пламен Матеев (И-т математика на БАН, Секция Вероятности и статистика, pmat@math.bas.bg) /

Заключение

WSD подходите показват засега способност да различават много добре главните значения на многозначните думи. Главните насоки на изследвания в момента са:

- начини за интеграция на всички възможни източници - семантични ресурси в статистическия подход към WSD;

- ускоряване на процеса на човешкото анотиране с максимално използване на автоматизирани процедури за извличане на значенията - с цел подготовка на учебни ресурси;

- усъвършенстване на самите алгоритми за WSD;

- прилагане на uns-WSD алгоритми в задачи, които не изискват съпоставяне на индуцираните значения със системните ресурси (главно речници). Такова приложение е IR, където засега WSD не допринася за съществено подобрение на извличането на документи.