Természetes nyelvek

Tartalom

- **Nyelvtechnológia elmélete**
  - **Nyelvtechnológiai alkalmazások**
  - **Morfológiai elemzés**
  - **Egyértelműsítés**
  - **Mondatelemzés**
  - **Szemantika**
  - **Szöveggenerálás**
  - **Diskurzus-reprezentáció**
- Számítógépes alkalmazások
  - Unifikációs nyelvtan
  - Statisztikai feldolgozás
  - Szövegkorpuszok
  - Lexikonok és szótárak
  - Szöveglétrehozás
  - Nyelvazonosítás
  - Számítógépes fordítás
- Egy intelligens webböngésző

Számítógépes alkalmazásai

- Nyelvhelyesség ellenőrzők
- Automatikus elválasztók
- Beszédfelismerők
- Szöveg-visszakeresők
- Automatikus szövegkivonatolók
- Számítógépes fordítók

Nyelvtechnológiai alkalmazások

- Nyelvfeldolgozás feladata: a szövegek aktuális nyelvi szintnek megfelelő gépi reprezentációja.
- Alkotóelemei:
  - formalizált nyelvtan :
    - lexikális rész
    - szabályrendszer
  - ezt kezelő program
- Megkülönböztetünk:
  - Morfológia vagy alaktan
  - Szintaxis vagy mondattan
  - Szemantika vagy jelentéstan
- Nem beszélhetünk emberi szinten történő szövegmegértésről.

Morfológiai elemzés

- Minimális egységek által hordozott információ -> szótár vagy lexikon:
  - Minden lehetséges szóalak megadása
  - Szótő + lehetséges toldalékok, képzők

- Problémát jelent a nem egyértelmű szótő és a toldalékok helyes kombinálása

Koskenniemi-féle kétszintes morfológia

- a formalizmus lexikonból és szabályokból áll.
- Lexikon = szótő + toldalék- >reguláris kifejezés
- I. Szint : nyelvi elemek lexikális reprezentációi

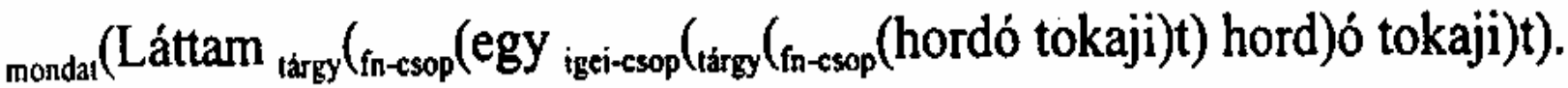
  autó + Ak + bAn
- II. Szint : szóalakok felszíni reprezentációi

  autó + 0k + ban
- A szabályok a két szint közti átmenetet definiálják

  Akkor és csak akkor nincs kötőhang a többes szám –k jele előtt, ha a tő utolsó hangja magánhangzó
- Hatékonyan implementálható
- Kétirányú = elemzés + generálás
- Helyesírás- ellenőrzésre használják

Egyértelműsítés

- A szavaknak többféle felbontása lehet
- Módszerei:
  - Szabályalapú: nem minden esetben használható, de ha igen kevés hibát követ el. Pl.: címkézet zárójelezés
    - A mondat nagy részének elemzését végzi
    - Példa:



  - Valószínűségi: minden esetben tud dönteni, de gyakrabban téved. Pl. HMM.

Mondatelemzés

- Felmerülő nehézségei:
  - Bonyolult szerkezetű mondatok
  - Ismeretlen szavak vagy szerkezetek
  - Egyértelműsítés
- Népszerű modell Noam Chomsky generatív grammatikája:
  - Definíció: G = (N,T,P,S)
    - N , T – nem terminális és terminális szimbólumok halmaza
    - P – szabályok, S – kezdőszimbólum
  - Végtelen sok mondat leírható egy véges szótár és egy szabályrendszer segítségével
  - Használt elemzési technika: LR(k) nyelvek

Mondatelemzés

- Problémát jelent a metaforikus szóhasználat
  - Megoldása lehetne a szótár nyitottá tétele
  - Ezáltal már nem fog tartozni Chomsky modelljéhez
  - Viszont csak bizonyos szófajú szavak kerülhetnek később a szótárba- > minimálnyelvtan, amit nem lehet nyitott osztállyal definiálni, és lesz a lexikonnak egy nyitott része.

Szemantika

- Szükség van atomi jelentésekre és ezek kombinálási szabályaira
- Alkalmazott formalizmusok:
  - Elsőrendű predikátumkalkulus
  - Montague nyelvtanok
  - szituációs szemantika (Barwise 1983)
  - frissítő szemantika
  - dinamikus szemantika
  - diskurzus reprezentációs elmélet

Keretszemantika

- Tudásreprezentációban használt kerethez hasonló információk alapján történik a feldolgozás
- Keret (frame): sztereotip szituációkat jellemző ismeretrendszer
  - vesz, elad, fizet, kerül stb.
  - pénz, fizetés, kereskedő, vásárló stb.
- Forgatókönyv (script): adott kerethez tartozó esemény részeseményeinek sorrendjét határozza meg
  - Senki nem tud venni, amíg valaki más nem akar eladni
- Fejlettebb modellek figyelembe veszik a szereplők céljait és a célok között fennálló viszonyokat
  - Péternek nagy adósságai voltak. Péter eladta a kocsiját.
  - Kérdés: Miért adta el Péter a kocsiját?
  - Válasz: Mert nagy adósságai voltak.

Szöveggenerálás

- Számítógépben tárolt ismeretek természetes nyelven történő megfogalmazása.
- Nehézségei:
  - a hosszabb koherens szövegek generálása, a létrehozás tervezési lépéseinek a kidolgozása.
  - lexikonbeli elemek helyes kiválasztása (szinonimák)
  - mondatok összefűzése, úgy hogy ne legyen köztük törés - >mondattervezés

Diskurzus-reprezentáció

- Kamp elmélete (1981):
  - minden D szöveghez tartozik egy diskurzus-reprezentáló szerkezet, amely D-t kvantormentes klóz-alakban ábrázolja
  - szöveg-reprezentációs szerkezet alakja: DRS=<REF, FELT>, ahol REF a DRS szövegreferenseinek, Felt pedig az egyedekre vonatkozó feltételeinek halmaza
  - a mondat rendszerbeli reprezentációja valamilyen DRS-ken operáló függvény lesz
- számítógéppel való ábrázolása:
  - a DRS egy állomány
  - diskurzusreprezentáció egy kártya.
- Legyen a következő diskurzus:

„András orvos. Ha egy orvosnak van számítógépe, akkor játszik vele.”

X1: { X1=András orvos(X1) }	X1, X2: { orvos(X1) számítógép(X2) birtokol(X1, X2) } => { játszik(X1, X2) }

Tartalom

- Nyelvtechnológia elmélete
  - Nyelvtechnológiai alkalmazások
  - Morfológiai elemzés
  - Egyértelműsítés
  - Mondatelemzés
  - Szemantika
  - Szöveggenerálás
  - Diskurzus-reprezentáció
- **Számítógépes alkalmazások**
  - **Unifikációs nyelvtan**
  - **Statisztikai feldolgozás**
  - **Szövegkorpuszok**
  - **Lexikonok és szótárak**
  - **Szöveglétrehozás**
  - **Nyelvazonosítás**
  - **Számítógépes fordítás**
- Egy intelligens webböngésző

Unifikációs nyelvtan

- Unifikációs formalizmusok:
  - Fejnyelvtan
  - Lexikális funkcionális nyelvtan
  - Fabővítő nyelvtan
  - Kategoriális unifikációs nyelvtan
- A nyelvi elemeket attribútum–érték párok halmazaként reprezentálják: jegy–érték együttesek
- Alulspecifikáltság: egy adott jegy jelen van, de értéke nem vagy csak részben meghatározott.
- Változókat is használhatunk
  - pl. alany és állítmány számának egyeztetésére
- Unifikáció = nyelvtani információk összeegyeztethetőségét vizsgálja

Statisztikai feldolgozás

- Nyelvfeldolgozás = információátvitel zajos csatornán
- A módszer alapelemei:
  - Átviteli modell = felismert kimenet valószínűsége
  - Nyelvmodell = egyes üzenetrészek adott környezetben való előfordulási valószínűségei.
- Legnépszerűbb alkalmazott modell a rejtett Markov modell (HMM)
  - a mondat szavai lesznek az észlelt állapotok
  - a szintaktikai osztályok (főnév, ige) a rejtett állapotok
  - a cél: a mondat minden egyes szavára a legvalószínűbb osztály megtalálása

Szövegkorpuszok

- Gépi nyelvfeldolgozás számára összegyűjtött szövegek együttese.
- Az egyes szavak különböző helyzetben való előfordulásainak tanulmányozására használják.
- Párhuzamos korpuszok = eredeti szöveg és a fordítása.
- Módszereire elsősorban valószínűségi és statisztikai módszerek jellemzőek
  - Pl. Olyan szerkezetekre alkalmazzák mint: erős légy

Lexikonok és szótárak

- Lexikális tudás = a nyelv szavainak, kifejezéseinek ismerete.
- Szótár = lexikális elemek listája + morfoszintaktikai, szemantikai, fonológiai viselkedésüket leíró jegyek összessége- >szükség van egy jegyleíró formalizmusra.
- A reprezentációs nyelv szabványosítása az SGML (Standard Generalized Markup Language), szótárak leírásához pedig a TEI (Text Encoding Initiative)- > formától függetlenül lekérdezhetővé válnak az egyes mezők és kombinációik.

Terminológiai adatbázisok

- Terminus = szakiránytól függő, akár teljesen más jelentéssel bír, állandóan születőben van.
- Terminológiai adatbázisok dinamikusak.
- Jellemzőek a soknyelvű adatbázisok.
- A fogalmak egy fogalmi hálózat megfelelő relációkkal elérhető csomópontjaként jelennek meg. Jellemzésük tezaurusz-deszkriptorokkal, szinonimákkal, rövidítésekkel, definíciókkal, képekkel, relációkkal stb. történik.

Szöveglétrehozás

- Szerzői eszközök: helyesírás ellenőrző, elválasztó, nyelvtani ellenőrző, szinonima szótárak.
- Hibák:
  - Billentyűzeten való melléütésből származó (környező betűk elhelyezkedése szerint)
  - Magyar – angol billentyűzeten való y z eltérés
  - Magyar ékezetes betűk szabványos vagy nem szabványos elhelyezése
  - Beszéd írásra való hatása „azt írjuk, amit mondunk”

Más műveletek

- Automatikus elválasztás
  - A szó minden lehetséges elemzését ismernünk kell. Pl. Legelőre
  - Az elválasztó úgy működjön, hogy jelenléte alig észrevehető legyen -> nem interaktív. A kézi elválasztás lehetősége biztosított kell legyen.
- Keresés:
  - Egy szó minden alakjának felismerése.
  - Probléma: a szavak nincsenek szótári alakban -> a mechanikus rendszerek gyakran tévednek.
- Nyelvhelyesség ellenőrző:
  - Egyenlőre csak szóellenőrzőkről beszélhetünk.

Más műveletek

- Szöveg visszakeresés:
  - Fontos a szinonimák illetve különböző nyelvekre történő fordítások közti keresés is
  - a szemantikát is figyelembe kell vennünk. Pl. Kutya – Kosárlabda EB
- Automatikus szövegkivonatolás:
  - Célja a szöveg tartalmának kevesebb mondatokkal való kifejezése.
  - Reális cél a szöveg releváns mondatainak kiemelése, és koherens szöveggé alakítása.
  - A kiválasztás statisztikai alapon vagy kulcsszavak alapján történik.

Nyelvazonosítás

- Feladat: a fordításnak a gépi, illetve géppel támogatott létrehozása, továbbá a forrás és a célszövegek szinkronizálása a későbbi feldolgozás számára.
- Elsősorban statisztikai alapon történik:
  - Nyelvek legrövidebb szavainak eloszlását figyelik
  - Egyes szó és karaktersorozatok gyakorisága
  - Nyelvre jellemző speciális karakter és karakterkombinációk megfigyelése.
  - Legelterjedtebbek a trigram-modellek, egymást követő betűhármasok gyakoriságainak megfigyelése.

Számítógépes fordítás

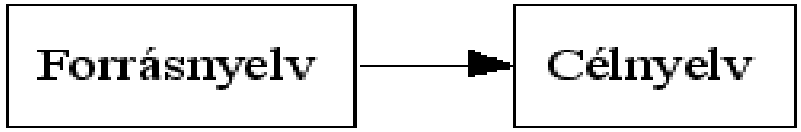
- Gépi fordításhoz használt számítógépes eszközök csoportosítása:
  - Teljesen automatizált gépi fordítás (TAGF)
    - Közvetlen emberi beavatkozás nélkül működő rendszerek
    - Legfeljebb technikai szövegek felszínes fordítására alkalmas.
  - Ember támogatta gépi fordítás (ETGF)
    - A gép a felhasználó segítségével ad választ a többértelműségekre és bizonytalanságokra.
  - Gép támogatta emberi fordítás (GTEF)
    - Hagyományos emberi fordítást jelent.
    - A fordító segédeszközei egy írógép és szótár funkcióját betöltő hatékony számítógépes rendszer.

Gépi fordítás csoportosítása

- Produktív:
  - a fordítás célnyelvén a mondatokat a program maga szintetizálja
  - technikai előkészítését a kontrollált nyelvi eszközök végzik.
  - Lehet:
    - Közvetlen, ha a forrásnyelv analízise és a célnyelv szintézise függő
    - Közvetett, ha független, ez továbbá lehet
      - interlingvális fordítás
      - transzfer fordítás
- Mintaalapú ha
  - csak kikeresi a forrásnyelv mondatai közül a leghasonlóbbat és annak „konzerv”-fordítását adja meg.
  - a forrásnyelv mondatait ún. fordítómemóriákban tárolják
  - elsősorban fordítómemóriákat és a velük társítható fejlesztéseket jelenti

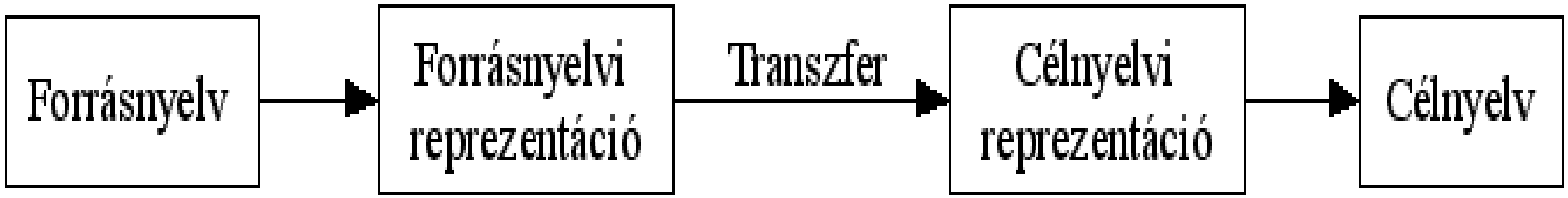
Produktív fordítási technikák

- Közvetlen fordítás:



- Közvetítő nyelv segítségével:

- Transzfer módszer:



Nem teljesen automatikus fordítás

- Felhasználó bevonása a fordítási folyamatba.
- Cél:
  - meglévő fordítások hatékony felhasználása
- Fordítói munkaállomások:
  - kétnyelvű szótárak
  - szaknyelvi terminológiai adatbázisok
  - fordítómemóriák
  - valódi gépi fordító rendszerek elérését is lehetővé teszik.

Tartalom

- Nyelvtechnológia elmélete
  - Nyelvtechnológiai alkalmazások
  - Morfológiai elemzés
  - Egyértelműsítés
  - Mondatelemzés
  - Szemantika
  - Szöveggenerálás
  - Diskurzus-reprezentáció
- Számítógépes alkalmazások
  - Unifikációs nyelvtan
  - Statisztikai feldolgozás
  - Szövegkorpuszok
  - Lexikonok és szótárak
  - Szöveglétrehozás
  - Nyelvazonosítás
  - Számítógépes fordítás
- **Egy intelligens webböngésző**

A LEXXE böngésző

- Elérése: www.lexxe.com
- Teszteljük a következő kérdésekkel:
  - Who did Bill Gates marry?
  - Who was killed by Lee Harvey Oswald?
  - Who is the best actress in the world?
  - Who assassinated President Lincoln?
  - How old is Yahoo?
    - A Google találatainak száma erre a kérdésre **64 000 000.**
- Hasonlítsuk össze a Google válaszaival ugyanezekre a kérdésekre?

Lexxe™ alpha

powered by advanced natural language technology

Please type your question or key words below >

How old is Yahoo? Find Answer Help

Answer: 10

Cluster: **Web:** **100 results found. Definitions: old Yahoo**

- it Seems Hard To Believe But Is Now 10 Years Old
- Ten Years Old Wednesday Still Searches the Internet
- Com: Is Ten Years Old Is Ten Years Old
- To Be 13 Years Old or over in Order To To Create a
- Nineteen Years Old
- 2
- Add Your News Search For Menopause at Years Old To My
- Add Your News Search For Years Old and Menopause To My
- Least 18 Years Old and That You Are the Legal

Yahoo! Netrospective: 10 years, 100 moments of the Web

Yahoo!'s picks of the top 100 moments from the first 10 years of the Internet. Inspired by 10x10 by Jonathan Harris.
http://birthday.yahoo.com/countdown - 8 KB

USATODAY.com - 10 years old and growing strong

... Ten years old Wednesday, Yahoo still searches the Internet ... Next month, Yahoo's re-energized entertainment division moves into the newly named Yahoo Center in Santa Monica ...
http://www.usatoday.com/money/industries/technology/2005-03-01-yahoo_x.htm - 57 KB

Yahoo! is ten years old (kottke.org)

is ten years old. posted March 03, 2005 at 10:13 am. Yahoo! turned ten years old this week which makes me feel like I'm about ready for the retirement home. As part of their celebration, they put up a copy of their home page from 1995.
http://www.kottke.org/05/03/yahoo-birthday - 4 KB

Yahoo! News Search Results for YEARS OLD AND MENOPAUSE

... Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]Search Home Help ... Add your news search for YEARS OLD AND MENOPAUSE to My Yahoo!: ...
http://news.search.yahoo.com/news/search?p=YEARS+OLD+AND+MENOPAUSE&ei=UTF-8 - 36 KB

Yahoo! News Search Results for menopause at years old

... Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]Search Home Help ... Add your news search for menopause at years old to My Yahoo!: ...
http://news.search.yahoo.com/news/search?p=menopause+at+years+old&ei=UTF-8 - 35 KB

Yahoo turns 10 years old March 2nd

Free ice cream! yahoo turns 10 years old march 2nd ... Home / Forums Index / Yahoo World / Yahoo Search. Yahoo Search ... message thread spans 2 pages: ([1] 2) > > Yahoo turns 10 years old March 2nd ...
http://www.webmasterworld.com/forum35/3193.htm - 20 KB

Wired 13.03: The UnGoogle (Yes, Yahoo!)

A LEXXE tulajdonságai

- Konkrét kérdésre konkrét választ ad, és a válaszon túl megjelenít néhány oldalt, ahonnan további információkat lehet kapni.
- Különböző típusú szövegeket tud azonosítani
- Szintaktikailag és szemantikailag elemez
- Képes a többértelműség feloldására a szövegkörnyezet felhasználásával.
- Képes megérteni a felhasználó szándékát és annak megfelelően válaszolni.
- Mindezek eredménye nagyon pontos és kielégítő válaszok.
- Szerzője egy számítógépes Enciklopédiát akart tervezni, amely kommunikál az emberrel egy adott témakörben.
- Harmadik generációs kereső ea LEXXE?

A LEXXE képességei

- A bemenetet nyelvként kezeli nem szimbólumokként
- Rendelkezik nyelvmegértő képességgel
- Nem képes szinonimák helyettesítésére
- A válaszadás statisztikus módszerek segítségével történik
- Megszorítás: a kérdések legfeljebb 10 szóból állhatnak, különben nem tudja jól elemezni azt
- Felismer lakcímeket, képzettségre és foglalkozásra vonatkozó információkat

Nyelvfeldolgozási elemei

- Amit a tervezői közzétettek:
  - Szövegfelismerő tehnológia
    - 20-40%-kal növeli meg a keresés hatékonyságát
  - Egy minimális dialóguskezelés is van benne, a kérdésekre adott válaszadás kapcsán
  - Klaszterezés (csoportosítás)
    - Nem hierarchikus csoportosítást használ, hanem a generált csoportok alapján újabb kereséseket végez.
- Ami még kell legyen benne
  - Szövegkivonatolás egyaránt kérdésből és a megtalált szövegekből

Lexxe™ alpha

powered by advanced natural language technology

Please type your question or key words below >

What is eigenvalue? Find Answer Help

Answer: **[1]** the variance in a set of variables explained by a factor or component, and denoted by lambda . An eigenvalue is the sum of squared values in the column of a factor matrix, or where a ik is the factor loading for variable i on factor k, and m is the number of variables. In matrix algebra the principal eigenvalues of a correlation matrix R are the roots of the characteristic equation **[2]** A scalar value that permits nonzero solutions y of equations of the form where L is a linear operator and where y can represent a vector or a function that is subject to certain boundary conditions. When y is a vector, L represents a matrix and y is termed an eigenvector. When y is a function, L can represent a differential or integral form, in which case y is called an eigenfunction.

Cluster: **Web:** **100 results found. Definitions: eigenvalue**

- Problems Adobe Pdf
- Problem Adobe Pdf
- Google Matrix
- Pdf]
- Eigenvectors
- Eigenfunction
- [adobe
- Matrices For the Symmetric
- Symmetric Matrices
- Largest
- Solution Methods
- Solution of Large
- Inverse
- Complex
- Solving
- Analysis

Eigenvalue -- from MathWorld
Eigenvalue -- from MathWorld Eigenvalue -- from MathWorld Eigenvalues are a special set of scalars associated with a linear system of equations (i.e., a matrix equation) that are sometimes also known as characteristic roots, proper values, or ...
http://mathworld.wolfram.com/Eigenvalue.html - 20 KB

Wikipedia: Eigenvalue
Wikipedia Free Encyclopedia's article on 'Eigenvalue'
http://en.wikipedia.org/wiki/Eigenvalue - 81 KB

eigenvalue
eigenvalue eigenvalue ... Let V be a vector space over a field k, and let A be an endomorphism of V (meaning a linear mapping of V into itself). A scalar ... is said to be an ... eigenvalue of A if there is a nonzero ... for which ... ...
http://planetmath.org/encyclopedia/Eigenvalue.html - 42 KB

V. - Thomas Pynchon
The biggest and most extensive Thomas Pynchon website on the Internet. Extensive guides to Gravitys Rainbow, V., Mason & Dixon, and more. ... The coefficient for their length is the eigenvalue. About 5 minutes of explanation with pictures will burn ... the tail end of this eigenvalue discussion, but I don't think it's ...
http://www.hyperarts.com/pynchon/v/extra/eigenvalue.html - 16 KB

eigenvalue -- Encyclopædia Britannica
eigenvalue one of a set of discrete values of a parameter, k, in an equation of the form Pψ = kψ, in which P is a linear operator (that is, a symbol denoting a linear operation to be performed), for which there are solutions satisfying ... an eigenfunction (proper or characteristic function) belonging to that eigenvalue. The totality of eigenvalues...
http://www.britannica.com/eb/article-9032125 - 40 KB

Eigenvalue, eigenvector and eigenspace

A LEXXE válaszadó rendszere

- A válasz generálásának lépései:
  - A kérdést átalakítja egy állítássá
  - Az állítás nem releváns elemeit elhagyja
  - Meghatározza a válasz legvalószínűbb témakörét. Információ-visszakereső rendszer segítségével megkeresi a témához kapcsolódó dokumentumokat
  - Ha valamelyik dokumentum valamelyik mondatának egy része illeszkedik a kapott állításra, azt találatként megjelöli
  - Nyelvészeti és statisztikai feldolgozás segítségével értelmes válasszá alakítja a találatokat.
  - Statisztikailag meghatározó szavak és szövegrészek alapján megadja a végső választ.

Összefoglaló

- Nyelvfeldolgozás elemei:
  - Szótár
  - Terminológiai adatbázis
  - Szöveg visszakeresők
  - Szemantikus elemző
  - Nyelvhelyesség ellenőrzők
  - Automatikus elválasztók
  - Nyelvazonosítók
  - Számítógépes fordítók
  - Automatikus szövegkivonatolók
  - Szöveggenerálók
  - Diskurzus reprezentálók
  - Beszédfelismerők

Végül **Hong Liang Qiao**, a LEXXE tervezőjének szavai

„In the next 5-10 years time, search engines will not be like Google and Yahoo today. They will just be something like LEXXE, a 3rd generation search engine, which are more intelligent and good at understanding human language.„