

Természetes Nyelvek Feldolgozása Modellei

Szófaji Címkezés Példaként

Liu Beier

QTZG9N

PTE TTK

2021. 05. 11.

Dr. Habil. Pauler Gábor felügyelet alatt

1. Bevezetés

Természetes nyelvek feldolgozása (Natural Language Processing; NLP), ami egy lényeges interdiszciplináris terület számítógéptudomány és nyelvészet között nagymennyiségű természetes nyelvi adatokat programmal elemzés értékében, egyre fontosabb szerepet játszik modern digitális világban gépi tanulási algoritmusokkal valószínűleg rengeteg alkalmazással, pl. automata keresési kulcsszó ajánlatban, idegennyelv fordítóban, nyelvészeti szakértő rendszerekben, intelligens személyi asszisztensnél, közösségi média felügyeletben és hirdetés testreszabásban stb. Az algoritmusai a gépi tanulási legkorszerűbb eredményeket alkalmazva sokféle sajátos fejlődéseket is mutattak a múlt év tizedekben, ezért a megvizsgálása kulcsfontosságú a mesterséges intelligencia ismertetése során.

NLP egy komplex terület sok alkérdéseket rendelkezve, amelyeket két szintre osztályozható, tehát alapvető feladatok és magasszintű feladatok. Az alapvető feladatokba tartoznak a szöveg szegmentálása (mondathatár detektálása, tokenizálás, morfológiai lebontás, kérdés-specifikus segmenálás), szófaji (Part-of-speech; POS) címkézés, és szintaktikai elemzés (chunking), pedig a magasszintű feladatokba a helyesírási/nyelvtani hibák felismerése és javítása, névelemek meghatározása (Named Entity Recognition; NER), szavak egyértelműsítése (Word Sense Disambiguation, WSD), tagadás és bizonytalanság felismerése, viszony kitermelése, stb. Alapvető feladatok inkább a nyelv formai szerkezetét vizsgálják, azonban magasszintű feladatok a nyelv jelentését is figyelembe veszik. E feladatok bonyolultsága és nehézsége nem csak az algoritmus bonyolultságától és összetettségétől függnek, mert a nyelvek maga is döntőek egyes szempontokból. Például a szófaji címkézési feladata viszonylag könnyebb agglutináló nyelveknek, hiszen a szóképzésekre és ragozásokra szolgáló morfémák száma és alakja meghatározottak, amelyek dektálásával a morfémákra vonatkozó szófaji csoportba könnyen címkézhetik a szóbanforgó szavakat. E feladat az izoláló nyelvek számára viszont sokkal bonyolultabb a kevesbé jellemezhető morfémák és alacsonyabb toldalék-tő arány miatt. Ezért a magyarnak is könnyebbnek tűnik a szófaji címkézése, amit következőképpen csak kevés magyar szakirodalom közvetlenül kezelte e feladatot. Ezzel szemben a mondatelemzési feladat nehezebb magyarban a diskurzus konfigurációs szórendje miatt (Kiss 1995), ami sok formális nyelvész figyelmét kelti, pl. az Alberti-Medve (2002) féle mondatelemzés modell az egyik legelfogadottabb megközelítés közoktatásban.

A szófaji címkézés relatív könnyűsége nem azt jelenti, hogy nem fontos magyaroknak a forgalom, főleg a vonatkozó modelllei és algoritmusai, ami segítség lesz Magyarország akadémikus globálizálódásának NLP szakkörben. A cikk arra szolgál, hogy a NLP főbb modelleket áttekinve, a NLP egyik legalapvetőbb feladatnak, tehát a szófaji címkézésnek fejlődését mutassa meg vonatkozó magyar szakirodalom bővítése érdekében.

2. Természetes nyelvek feldolgozása

Az NLP az 1950-es években kezdődött, mint a mesterséges intelligencia és nyelvészet területek metszése. Az NLP eredetileg különböztette a szövegnek az információ-visszakeresésétől (Information Retrieval; IR), amely rendkívül skálázható statisztikai alapú technikákat alkalmazva nagy mennyiségű szövegek indexelését és keresését végzi hatékonyan (Manning et al. 2008). Idővel azonban az NLP és az IR valamennyiben konvergáltak. Jelenleg az NLP több elég változatos területről tudást igényel a mostani NLP kutatókat mentálitásváltási és tudástárbővítési szükségekkel terhelve. E fejezet NLP fejlődését történetileg megvizsgálva, a főbb modelleket össze próbálja foglalni.

2.1. Szabály-alapú megközelítés

A korán leegyszerűsített megközelítések, pl. egy-egy szó szerinti gépi fordítás nyelvek között (Hutchins 1954), homográfiák és stilisztikai eszközök (pl. metafora) miatt el lett veszve. Utána Chomsky (1956) formálizmusi elméleti elemzése nyelvtanról becslést adott a problémák nehézségének, amely befolyásolta a Backus-Naur-Forma (BNF) azaz egy „kontextusmentes nyelvtan” (Context-free Grammar; CFG) létrehozását 1963-ban, ami gyakran a programozási nyelv szintaxisának képviselőjére szolgál. A nyelv BNF-s specifikációja egy derivációs szabályok halmaza, amely együttesen ellenőrzi a programkódot szintaktikailag. (A „szabályok” itt abszolút korlátok, nem a szakértői rendszerek heurisztikája.) Chomsky (1959) is felismerte egy még korlátozóbb „reguláris” nyelvtanát, amire adódott a Reguláris kifejezések (Regular Expression; RE; Kleene 1956), ami először a Ken Thompson *grep* segédprogramja támogatta UNIX rendszerben (Kernighan 1984). Ezt követően, 1970-es években, lexikális elemző (lexer) és mondattani elemző (parser) generátorok, pl. a *lex/yacc* kombináció (Levine et al. 1992) hasznosította a szóhajó nyelvtanokat. Egy lexer átalakítja szövegeket tokenekké és egy parser érvényesít egy tokeni szekvenciát. A lexer/parser generátorok nagy mértékben leegyszerűsítte a programozási nyelvek végrehajtását az RE és BNF specifikációit bemenetként véve, kódot és keresőtáblákat döntéshozatért generálva. Míg a CFG-k elméletileg nem elégedők igazi természetes nyelveknek, gyakran alkalmazzák őket NLP-s gyakorlatban. A végrehajtása egyszerűsítése érdekében a programozási nyelveket általában szándékosan tervezték egy korlátozó CFG változatával, azaz egy LALR (1) parserral (Előrelátó parser balról-jobbra terjeszkedő feldolgozással és alulról-felfelé irányuló derivációval; Aho 2007). Aztán kitalálták a felülről lefelé irányuló Prolog nyelvet is (1970, amely könnyebben végrehajtható de lassabban mint az alulról felfelé párja. Szintaxisa különösen alkalmas nyelvtanok írására, bár a szabályokat másképp kell megfogalmazni (azaz jobb-rekurzív módon) mint a *yacc* stílusú elemzőnek szántak. Ezek a kézzel írt korlátozott szabályok rugalmas bizonytalan természetes nyelvekkel szemben két fő problémát találtak: (1) Az ilyen NLP-nek jelentést (szemantikát) szövegből végig kellett kibontania; (2) A kézzel írt szabályok beszélt prózát és terület specifikus kontextust rosszul kezelte meg.

2.2. Statisztikai elvű megközelítés

A 80-as években egyre fejlődött valószínűségi elméletek befolyásolása alatt NPL területen megjelent egy fontos módszerváltás, ahol az egyszerű és markos közelítésekkel helyettesítették a mély és makacs szabályokat gépi tanulásban felhasznált statisztikai modellek segítségével, hogy jegyzetekkel ellátott korpusos rendszereket hoztak létre tömeges nyelvi adatokkal szigorúbb standardok alatt e modelleket tanítani. (Részletes összefoglalásért: Klein 2010) A mai mély tanulásig rengeteg modellt alapítottak NLP-ért, és legtöbbjük kevesebb általánosított szabályokkal leghatékosabb eseteket kitalálva értelmezést végzik felügyelt ill. felügyeletlen állapotban.

E gépi tanulási modellek nagyjából két kategóriába sorolhatók: tehát generatív vagy diszkriminatív. A generatív módszerek valószínűségeloszlásban gazdag modellek létrehozására törekcszenek így szintetikus adatokat is lehet „generálni”, tehát a generatív osztályozók megtanulják az bemenetek (x) és címkék (y) együttes valószínűségét ($p(x, y)$) és megjósolják a Bayes-szabályok használatával a $p(y|x)$ kiszámítását, majd kiválasztja a legvalószínűbb y címkét. Azonban a diszkriminatív osztályozók közvetlenül az posteriori valószínűség $p(y|x)$ -t modellezik, vagy közvetlenül az x bemenetekről az osztálycímkekre leképezést építenek. Generatív modellek aszimptotikus hibájára gyorsabban közelít tanuló halmaz bővülésével, azonban nagyobb mint a diszkriminatív modelleké, így e kettő módszer egymást kompenzáló. (Ng & Jordan 2001) Hasonlóképpen a generatív modellekhez képest, amelyek megoldhatatlanná válhatnak sok sajátosság felhasználásakor, diszkriminatív modellek általában több sajátosság használatát teszik lehetővé (Elkan 2012).

Minden statisztikai modellnek saját előnyei is korlátozásai vannak, hogy a fogalmuk és tulajdonságaik megvizsgálása nagy mértékben segítheti a kutatók szakszerűsítését. Itt található néhány gyakoribban használt modell:

2.2.1. HMM

A Rejtett Markov Modell (Hidden Markov Model; HMM) olyan generatív rendszer, ahol egy változó változhat (változó valószínűségekkel) több állapot között, a több közül egyik lehetséges kimeneti szimbólumokat generálva minden változással (változó valószínűségekkel). A lehetséges állapotok és egyedi szimbólumok halmaza nagyok lehetnek, de véges és ismert. Megfigyelhetjük a modell kimeneteit, de a rendszer belső elemei (azaz az állapotváltozás valószínűségei és a kimenet valószínűségei) „rejtve vannak”. Könnyebben meghatározva a HMM egy Markov láncban végző folyamat, amelyben a belső valószínűsége rejtett, pedig a Markov lánc egy valószínűséggel bíró véges állapotú gép (Finite State Machine; FSM), ahol a valószínűsége $\Pr(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \dots)$, $1 \leq i, j, k \dots \leq N$ (*) feltételes valószínűséggel jelemmez. (*) érvényesüléshez a HMM két egyszerűsítő feltételt használ, amely igaznak tűnt számos valódi jelenségben: (1) Az új állapot (vagy ugyanaz az állapot) az előző N állapottól függ. A legegyszerűbben „elsőrendű” esetén, ez a valószínűség csak a jelenlegi és előbbi állapotoktól függ: $a_{ij} = \Pr(q_t = S_j | q_{t-1} = S_i)$, $1 \leq i, j \leq N$. (2) Egy

állapotnak a kimenetei generálási valószínűsége csak az állapot önmagától függ.

Három alapvető kérdésre bomlik a modell Jack Ferguson szerint: (1) Becslés: adott kimeneti szimbólumok bizonyos sorrendje alapján egy vagy több lehetséges állapotváltozási valószínűségek szekvenciája kiszámítása (Adott HMM, S (állapotok), x (kimeneti szekvencia); Keressen: $\Pr(x|S)$); (2) Mintaillesztés: egy adott kimeneti-szimbólum sorrend létrehozására legvalószínűbb állapotváltozási sorrend megkeresése (Adott HMM, S , x ; Keressen: Q (állapotváltozási szekvencia), amely maximalizálja $\Pr(x,Q|S)$); (3) Tanulás: a kimenet-szimbólum sorrenddel rendelkező (tanító) adathalmaz megadásával, a legjobban beillesztő állapotválás/kimenet (belső) valószínűségei kiszámítása (Adott HMM (ismertelen), S , x ; Keressen: a (belső valószínűségek), amely maximalizálja $\Pr(x|a,S)$). (2) és (3) kérdés valójában a szekvenciára kiterjesztett naiv Bayes-i érvelés. (Bővebb a kérdésekről és alkalmazásáról: Lawrence 1989) A kérdések megoldásához számos algoritmusok léteznek, amely közül a leghíresebbek az előre-hátra (Forward-Backward) algoritmus, a Baum-Welch algoritmus, ami egyfajta elvárást maximalizáló (EM) algoritmus, és a Viterbi algoritmus. (Algoritmusok részletei: Viterbi 2009; Welch 2003; Baum et al. 1970; Dempster et al. 1977)

2.2.2. MEMM

HMM modellek szófaji címkézésben, szöveg szegmentálásában, NEM-ben, és információ kitermelésében találta helyét, viszont kettő problémája maradt: (1) HMM második feltételétől tévedhet mert gyakorlatban megfigyelések sajátosságai számtalanok és egymástól függő (2) a HMM valószínűség maximalizálásával egy generatív együttes modellt épít, de a megoldandó kérdések lehetnek feltételes adott megfigyeléssel. Ezeket specifikusan kezelve McCallum et al. (2000) egyféle információ elmélettel együtthető Maximumális Entrópia Markov Modelleket (MEMMs) hoztak létre, amely egy exponenciális feltételes (diszkriminatív) modell általánosított iteratív skálázással (Generalized Iterative Scaling; GIS; Darroch és Ratcliff 1972) tanított EM algoritmussal hasonló bonyolultsággal. MEMM a HMM bonyolult belső valószínűségi paramétereit helyettesítette $\Pr_{s'}(s|o)=\Pr(s|s',o)$ átmeneti függvényrel, ahol az s' az előző állott és az o a jelenlegi megfigyelési adat, ami kizárólag jelenlegi állapottól függő kimenet helyett több megtörtént állapottól függő állapotváltozást jelölő bemenetté vált. Így a MEMM paramétereit becselő algoritmust leírható: (1) Bemenetek: Megfigyelési adatszekvencia $o_{1:m}$, vonatkozó címkesorozat $l_{1:m}$, vonatkozó címkével és potenciális korlátozott átmeneti szerkezettel rendelkező állapotok $s_{1:N}$, és megfigyelés-állapot sajátosságai; (2) A megfigyelés-címke szekvenciához társított állapotszekvencia meghatározása (bizonytalan helyzetben két-lépésű EM iterációval); (3) A modell tanítása mindegyik előző átmeneti függvénynek egy jelenlegi állapot-megfigyelés pár (s,o) hozzáadásával; (4) A maximális entrópiás megoldás keresése minden diszkriminatív függvényhez GIS-n keresztül; (5) Kimenetek: Maximum entrópia alapú Markov modell, amely egy felcímkézetlen megfigyelési szekvenciának előrejelzi a megfelelő címkéjeit.

2.2.3. CRF

Lafferty vezetése alatt, McCallum és Pereira egyre jobban fejlesztették a diszkriminatív megközelítésüket egy ún. Feltételes Véletlen Test (Conditional Random Field; CRF) modellre 2001-ben, amelyet több alkalomra is alkalmaztak (Culotta et al. 2005; Sutton és McCallum 2004). A CRF-ek általánosítják a logisztikai regressziót sorozatos adatokra ugyanúgy, ahogy a HMM-ek általánosítják a Naive Bayes-t.

A Markov féle modellek egyik fő problémája a címke torzítása, ahol az átmeneti pontszámok (valószínűségek) súlyát a következő állapotokba el kell osztaniuk, viszont a megfigyelési adat nem befolyásolhatja egyes állapotokban súlyai nagyságát, ezért a Markov modellek a kevesebb követő állapottal rendelkező állapotot torzítik. (Bottou 1991) CRF ezt a problémát kezelte azzal, hogy MEMM minden állapot feltételes valószínűségéhez rendelt exponenciális modellek helyett CRF egyetlen exponenciális modellt ad az egész címkeszekvencia együttes valószínűségének adott megfigyelési adatszekvenciával.

2.2.4. N-gram

N-gram egy magasrendű Markov féle modell, amely különösen NLP-ra fókuszál automatikus szó/kifejezés kiegészítése, helyesírási javítás, beszédek felismerése, ill. szó egyérterműsítése, stb. magasszintű feladatokban. Nylevekben egyes szavak/kifejezések bizonyos szerkezetben, ill. kontextusban gyakrabban fordulnak elő, így a nyelvi adatszekvenciák valószínűsége kiszámításával, a legvalószínűbb értelmezését kitalálható. pl. instrumentális eset –vAl 'v' betűje mássalhangzóval végződő szavaknál mássalhangzóvá válik, így a nagy előfordulási gyakoriságot tekintve, N-gram könnyen helytelen írásokat megtalál és javít, ill. INSTR esetet kijelöl.

2.2.5. SVM

A támogatási vektoros gép (Support Vector Machine; SVM) egy másik NLP-ben gyakran alkalmazott diszkriminatív modell, amely egy hipersíkkal nyelvi adatokat külön csoportba (pl. szavakat szófajokba, főneveket esetekbe, stb.) beoszt, hogy az adatpontoktól síkig távolsága minél nagyobb legyen. Így ha az adatok lineáris választható két osztályba egy $(w \cdot x) + b = 0$ hipersíktól távolság maximizálása ugyanaz mint $\min \frac{\|w\|^2}{2}$ megkeresése, ahol $y_i[(w \cdot x_i) + b] - 1 \geq 0$, $y_i = \text{sgn}[(w \cdot x_i) + b]$, $(i = 1, 2, \dots, n)$, amelyet Lagrange függvényrel oldható; ha lineárisan nem választható adatokról szól, egy $K(x, u) = \langle \varphi(x), \varphi(u) \rangle$ kernel függvényrel, amely pozitívdefinit és a Mercer (1909) tételét kielégíti, az adatokat magasabb dimenzióba átalakíthatják, hogy ott választható legyen. (Boser et al. 1992) Többosztályos SVM modellek pedig csak 20. század végén jött létre amikor általánosították a hipersíkos megközelítést k mennyiségre, hogy a döntési függvény $f(x) = \arg \max_n [w_n \cdot x + b_n]$, $n = 1, 2, \dots, k$ legyen (Weston és Watkins 1999) ill. a problémát irányított körmentes gráffal megközelíse (Directed acyclic graph; DAG; Platt et al. 2000). (Bővebb: Awad és Khanna 2015)

3. Szófaji címkézés

Szófaji címkézés (Part-of-speech tagging; POST) az egyik legalapvetőbb feladat korpus nyelvészetben ill. NLP-ben, ahol a tokenizált szavakat a morfoszintaktikai szerepköre szempontjából címkézik egy adott címkehalmazzal, pl. a magyar nyelv két hazánkban és nemzetközien ismert címzehalmazzal rendelkezik, azaz emMorph kódlista (e-magyar.hu 2015), és a MSD kódrendszer (Nyelvtudomány Kutatóközpont 2014), de egy un. HunMorph annotációs rendszer (Rosen 2006) is létezik. E feladat főbb hagyományos és korszerű teljesítményekről ír a fejezet.

3.1. Hagományos modellek

POST legkorábbi modellek a NLP kezdetével egyezik, a szabály-alapú modelltől statisztikai elvű modellre átalakulása jellemző, itt egyik legelfogadottabb szabály-alapú modellről, négy statisztikai modellről, és egy alternatív tanulásról lesz szó.

3.1.1. TBL

Transzformáció-alapú tanuló (Transformation-based learning; TBL) NPL modellt Brill (1995) írta le transzformációs szabályok alapján egy statisztikai modellekhez képest egyszerűbb értelmezésért. Az algoritmus a kezdeti helyzetben annotált szöveget sokszor átalakítja, és mind egyszer a lehetséges átalakítások közül azt választja, amely legnagyobb hibacsökkentést ad. A transzmormáció két elemből áll, azaz egy újraíró szabály, ami egyik token szófaját változtatja, és egy felidéző környezet, ahol célszerű újraírást működtetni, pl. „enni adat lónak” szókörnyezetben az „adat” nem főnév, hanem ige és ha hiba történt első annotációnál, kézzel írott újraíró szabállyal javítható. A POST feladatban e X szófajról Y szófajra újraíró folyamat történik akkor és csak akkor, ha (1) a szó korpusban nem található, vagy (2) a szó legalább egyszer Y szófajjal megjelent korpusban.

3.1.2. TnT

Trigram'n'Tags (TnT) egy másodikrendű HMM alapú POST modell Brants (2000) által jött létre 96,7% (ismertelen szavak esetén 85,5-89%) pontossággal Penn Treebank és NEGRA korpus tokenjait címkézve. A címkehalmaz $t_{-1:T+1}$ állapotok, ahol t_{-1} , és t_{T+1} szekvencia kezdete és vége, és $w_{1:T}$ kimenetek, így a célfüggvény $\arg \max_{t_{1:T}} [\prod_{i=1}^T \Pr(t_i | t_{i-1}, t_{i-2}) \Pr(w_i | t_i)] \Pr(t_{T+1} | t_T)$. És egy unigrammal, bigrammal, és trigrammal közelített $\Pr(t_3 | t_1, t_2) = \lambda_1 \widehat{\Pr}(t_3) + \lambda_2 \widehat{\Pr}(t_3 | t_2) + \lambda_3 \widehat{\Pr}(t_3 | t_1, t_2)$, $\lambda_1 + \lambda_2 + \lambda_3 = 1$ interpolációval ritka adatalmazon tanította. A modell ismertelen szavakat és nagybetűs írást is kezelt. A munka azt mutatta, hogy a Markov modellek legalább olyan hatékony lehet mint a Maximális Entrópiás modellek.

3.1.3. MEM

Habár TnT területi mérföldkőként legyőzte a Ratnaparkhi (1998) Maximális Entrópia modelljét (MEM), egy tudásforrást gazdagító MEM újabb verzió (Toutanova és Manning 2000) hamarosan kihívást feltett. A modell arra figyelte, hogy az előző megközelítések semelyik sem bővítette az információforrási csatornáját, ami szintén korlátozó, ezért ők egy kontextustól információt kitermelő modellt írtak. E MEM-nek egy alapvonalas változatával rendelkezett, amely viszont nem adott kitűnő teljesítményt (96,72% átfogó pontossággal kicsivel megnyerte a Ratnaparkhi MEM-jét alacsony ismertlen szavakat címkéző képességgel). De továbbra három különleges sajátosságot, mármint nagybetűs írást, igei formákat, és partikulákat, egymás mellett alapvonalas modellhez hozzáadva jobb eredményeket sikerült kihozni (96,86%/86,91%).

3.1.4. Stanford címkéző

A Stanfordi Tudományegyetem kutatócsapat (Toutanova et al. 2003) az MEM felül folytatták a munkáját egy ciklus függőségi hálózat (Cyclic Dependency, Network; CDN) koncepcióval címkézési pontosságát 97,24%/89,04%-ra növelve. A modell legjellemzőbb eltérés egy csupa Bayes hálójától az, hogy bizonyos állapotok függőségei kétirányúak, így az együttes valószínűségük feltételes valószínűségi formulával nem kifejezhető. E problémát egy pontszám maximalizálásával megoldták, amely egy $score(x) = \prod_i P(x_i | Pa(x_i))$ ($Pa(x_i)$ az x_i -re vezető állapotok halmaza) lokális valószínűségek sorozata és körmentes helyzetben a pontszám együttes valószínűségi formulát ad. Manning (2011) ezt a címkéző teljesítményét 97,32%/90,79%-re javított, és rendszeresen elemezte a lemaradt hátrányokat, amelyek a modellt 100%-tól megakadályozták, pl. Lexikai ritkaság (Asztalos több esetben NN de JJ-ként ritkán fordul elő), ismertlen szavak, kétértelműség kontextus hiányában (Elég (VERB vagy ADV) az asztalon lévő papír), arany szabály inkonzisztens, hiányos, vagy hámis, stb.

3.1.5. SVMTool

Az MEM-féle megközelítés népszerűsülése során, SVM modellek is saját utat találtak Spanyolországban. Giménez és Márquez (2004) egy C és C++ nyelvekkel alkalmazott SVMTool névű címkéző és címkézőt generáló szoftvert készítettek, amely az egyszerűsége, rugalmassága, markossága, hordozhatósága, pontossága, és hatékonysága miatt hírnevet nyert. Az SVMTool praktikusság orientált, végrehajtási sebességéért kicsivel elengedheti már a pontosságát, amely angol Wall Street Journal (WSJ) korpus esetén 97,39%/89,01%, és spanyol LEXESP korpus esetén 98,08%/88,28%. E szoftver három komponenssel rendelkezik, tehát (1) SVMTlearn, amely SVM-Light felhasználásával adott adatállományok megtanulásával SVM modellt készít; (2) SVMTagger, amely szövegtár és tanult SVM modell alapján címkézési feladatot végez; (3) SVMTeval, amely a megcímkézett kimenetek adott standarddal összehasonlítva modellbecslést ad.

3.1.6. MBT

Egy korábban fogalmazott felügyelt induktív memória-alapú címkéző (Memory-Based Tagger; MBT; Daelemans et al. 1996), amely kivitelesen nem modell-alapú, inkrementálisan fogadja be címkével rendelkező tanuló adatokat, és memóriájába ezeket a tapasztalatokat tárolja. Ismertelen teszt adatoknak mindig a legközelebb tanított adatpéldát felidéz, amely címkéjét osztályozandó teszt adatnak hozzáadja.

3.2. Korszerű modellek

Collobert et al. (2011) a hagyományos NLP feladatot átfogón elemzte amely szintén egy úttörő a terület korszerűsítésére. Mély tanulási modellekkel, nem statisztikai megközelítésű modellekkel, és terület-specifikus modellekkel, a mostani POST feladat változatosabbá, kreatívabbá, és könnyebbé válik. alább néhány legjellemzőbb trendet, ill. értékes munkát mutat.

3.2.1. Többszínű tartalmak

Az elemzendő nyelvadatok tartalma gazdagabbá válik. Egyrészt NLP-t egyre jobban részletes szakterületre alkalmaztak, amely új kérdéseket indokol. Rengeteg latin-görög eredetű szakkifejezéssel (Deléger et al. 2009) és mozaikszavakkal (Nadeau és Turney 2005) rendelkezve, klinikai nyelvadatok különleges morfoszintaktikai elemzést és NER-t igényel, hogy velük emelkedett az „ismertelen” címkét bíró tokenok száma is. Coden (2005 et al.) már feltaláltak egy terület-specifikus POST modellt MED korpust felhasználva, Barrett és Weber-Jahnke (2014) egy token-középpontú címkézővel (Token-Centric-Tagger; TcT), amely egy nem-dinamikus programozási modell naív-Bayes osztályozóval, a élettan-orvossági szövegekelemző pontosságot 83,9%-ra emelt (itt látható az általános szöveghez képest pontosság csökkenést). Üzleti területre is kezdtek szálni különleges erőt (Danenas et al. 2020), hogy az OMG (2017) által meghatározott Üzleti Szókincsek és Szabályok Szemantikájából (Sementaics of Business Vocabulary and Business Rules; SBVR) információt kivonják, viszont a nyelvtani címkéző a használati eset modellel zavarosan nem egyezik sok esetben. Másrészt a stilisztikailag sokszínű helyzeteket is próbálják elemezni, elsősorban a társalgási stílust közösségi médiáktól kitalált adatokkal. Twitter as első olyan online közösség, amelyet NPL modellel feltártak. Gimpel et al. (2011) egy hiteles POST munkát végez ezen, hogy egy Stanford címkéző (85,85%)-nél sokkal erősebb 89,37% átfogó teljesítményt ért el. A modell CRF-alapú, twitter különleges jelhalmazhoz készített kódlistával rendelkezik, amely egy vázlatot kínált az utóbbi fejlesztések számára, pl. Beißwenger et al. (2015) az un. STTP német kódlistát számítógép-közvetített kommunikációban jellemző jelenségekkel (pl. E: emotikon :))) bővítették.

3.2.2. Többféle nyelv

Ezenkívül, a POST feladatot különböző nyelveken tesztelték a hagyományos időszakban kizárólagos angol korpust elemző helyzetet leküzdve. Arab (Alhasan és Al-Taani 2018), vietnami (Ngo et al. 2017), különböző kisnyelvek Indiában, és még kisebb nyelvek (Alex és Zakaria 2013; Ptaszynski és Momouchi 2012) is a nyelv sajátosságát észre véve, változatos POST modelleket találnak fel.

3.2.3. Többszintű reprezentáció

Mély neurális hálózatok (Deep Neural Network; DNN) egy féle legnépszerűbb mély tanulási algoritmus, amely több egymástól szemantikailag eltérő reprezentációs szinttel alakul ki. E gondolatot követően, Dos Santos és Zadrozny (2014) CharWNN modellt írtak le, amely egy különösen POST feladatra készített DNN változata egy konvolúciós réteggel kerekter-szintű reprezentációt kitermelni. CharWNN első reprezentációs szintjén egyes szavakat valós értékű sajtvektorokra átalakít lineáris traszformáció

$$r^{wrd} = W^{wrd} v^w, r^{wrd} \in \mathbb{R}^{d^{wrd}}, W^{wrd} \in \mathbb{R}^{d^{wrd} \times |V^{wrd}|}, v_i^w = \begin{cases} 0, & i \neq w \\ 1, & i = w \end{cases}, i = 1, 2, \dots, |V^{wrd}|$$

által. A kerekter beágyazása r^{char} hasonló eljárást végez. A konvolúciós réteg először egy k^{char} méretű vektort képez, ahol az m -dik kerekter beágyazása a vektor középpontja. Utána $[r^{wrd}]_j = \max_{1 \leq m \leq M} [W^0 z_m + b^0]_j$ tanulandó traszformációval N

darab szó beágyazását is képezhetik. Colbert et al. (2011) féle pontszámfüggvény értéke maximalizálásával, e modellt sztochasztikus gradiens-ereszkedésű tanításra elkészíthetik. E modell WSJ korpust 97,32% és portugál Mac-Morpho korpust 97,47% átfogó pontossággal címkézhetette. A hagyományos neurális hálózatnak a tanító modelljét két új logaritmus lineáris modellre, tehát folyamatos szócsoomag (Continuous Bag-of-Words; CBOW) modellre, ill. folyamatos ugró-gram (Continuous Skip-gram) modellre is fejlesztik, hogy csökkenjen a bonyolultsága a szavak reprezentációs (Word Representation; WR) modellnek (Mikolov et al. 2013).

Az un. kétirányú hosszú rövid-távú memória (bidirectional long short-term memory; bi-LSTM; Plank et al. 2016) egy másik megközelítés mély neurális hálózathoz korábban tárgyalt kétirányú rekurrens neurális hálózatról (bidirectional recurrent neural network; bi-RNN) fejlődve. Egy bi-RNN lehet szkvenciás: $v = bi - RNN_{seq}(x_{1:n}) = RNN_f(x_{1:n}) \circ RNN_r(x_{n:1})$ vagy kontextusos: $v_i = bi - RNN_{ctx}(x_{1:n}, i) = RNN_f(x_{1:i}) \circ RNN_r(x_{n:i})$, ahol (\circ) egy konkatenáció oda (f) és vissza (r) irányuló RNN-ek között. Bi-LSTM egy kontextus bi-RNN és LSTM (Hochreiter és Schmidhuber 1997) kombinációja, amely 22 nyelven POST feladatot végzett először 96,50% átlagos teljesítménnyel. A többnyelvű POST munkát a fenti említett két modell ösztéte is megvalósította (Pota et al. 2018), ahol beágyazott rétegek, bi-LSTM rétegek, és teljsen kapcsolt rétegek egymást követve egy mély szerkezetek megőriztek olasz 98,8%, angol 94,9% teljesítménnyel.

4. Összefoglalás

A természetes nyelvfeldolgozás egy jelentőségteljes mesterséges intelligenciai munkakör, amely lényegesen megtanulásához folyamatos kitűzés és logikus gondolatfolyamat szükséges. A megértését biztosító első lépésekbe tartoznak a nyelvtudomány és számítógéptudomány részletesebb ismeretei, az NLP alapvető feladatokról, különösen a szófaji címkezés feladatokról konszistens kutatás, és a gépi, ill. mély tanulás főbb modellei és algoritmusai hajlékony manipulációja. A POST egyre jobban bekeveredett többi magasabb szintű feladatokba, amelyek csővezetéküket (pl. GATE (Cunningham et al. 2011) megőrizve holisztikusan egyre bonyolultabb szituációkat kezelhetnek. A legkorszerűbb modellek általában Markov- és DNN-féle, de felügyelt ill. nem statisztikai pl. evolúció-alapú (Forsati és Shamsfard 2014) modellek is saját jövőjüket megtalálják.

Hivatkozások

- Aho, A. V., Lam, M. S., Sethi, R., & Ullman, J. D. (2007). *Compilers: Principles, Techniques, Tools* (2nd. edn.). Reading, MA: Addison-Wesley. [http://ce.sharif.edu/courses/94-95/1/ce414-2/resources/root/Text%20Books/Compiler%20Design/Alfred%20V.%20Aho,%20Monica%20S.%20Lam,%20Ravi%20Sethi,%20Jeffrey%20D.%20Ullman-Compilers%20-%20Principles,%20Techniques,%20and%20Tools-Pearson_Addison%20Wesley%20\(2006\).pdf](http://ce.sharif.edu/courses/94-95/1/ce414-2/resources/root/Text%20Books/Compiler%20Design/Alfred%20V.%20Aho,%20Monica%20S.%20Lam,%20Ravi%20Sethi,%20Jeffrey%20D.%20Ullman-Compilers%20-%20Principles,%20Techniques,%20and%20Tools-Pearson_Addison%20Wesley%20(2006).pdf)
- Alberti, G., & Medve, A. (2002). *Generatív grammatikai gyakorlókönyv*. Janus/Books (Gondolat), Budapest; 2 (javított) kiadás: 2005.
- Alex, M., & Zakaria, L. Q. (2013). Kadazan Part of Speech Tagging using Transformation-Based Approach. In: *The 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013)*, Procedia Technology, 11 (2013), 621 – 627. <https://doi.org/10.1016/j.protcy.2013.12.237>.
- Alhasan, A., & Al-Taani, A. T. (2018). POS Tagging for Arabic Text Using Bee Colony Algorithm. In: *The 4th International Conference on Arabic Computational Linguistics (ACLing'18)*. Procedia Computer Science 142 (2018), 158–165. <https://doi.org/10.1016/j.procs.2018.10.471>.
- Awad, M. & Khanna, R. (2015). Support Vector Machines for Classification. In: Khanna, R. & Awad, M. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress, 39-66. https://www.researchgate.net/profile/Mariette-Awad/publication/300723807_Support_Vector_Machines_for_Classification/links/5b5ed9b1458515c4b25271f7/Support-Vector-Machines-for-Classification.pdf?origin=publication_detail
- Ngo, B. X., Nguyen, L. D., & Tu, P., M. (2017). An empirical study on POS tagging for Vietnamese social media text. *Computer Speech & Language*, 50, 1-15. <http://dx.doi.org/10.1016/j.csl.2017.12.004>
- Barrett, N., & Weber-Jahnke, J. (2014). A token centric part-of-speech tagger for biomedical text. *Artificial Intelligence in Medicine*, 61(1), 11-20. <https://doi.org/10.1016/j.artmed.2014.03.005>.
- Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1), 64-171. Retrieved May 12, 2021, from <http://www.jstor.org/stable/2239727>
- Beißwenger, M., Bartz, T., Storrer, A., & Westpfahl, S. (2015). Tagset and guidelines for the PoS tagging of language data from genres of computermediated communication / social media. <https://core.ac.uk/download/pdf/83654472.pdf> (accessed 13 May 2021).
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992) A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92)*, Pittsburgh, 27-29 (July 1992), 144-152. <https://www.svms.org/training/BOGV92.pdf>
- Bottou, L. (1991). *Une approche th'eorique de l'apprentissage connexionniste: Applications `a la reconnaissance de la parole*. Doctoral dissertation, Universit'e de Paris XI. <https://leon.bottou.org/publications/pdf/bottou-1991.pdf>
- Brants, T. (2000) TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of the sixth conference on Applied natural language processing (ANLC '00)*. Association for Computational Linguistics, USA, 224–231. <https://doi.org/10.3115/974147.974178>
- Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computer Linguistics*, 21(4), 1-37. <https://www.aclweb.org/anthology/J95-4004.pdf>
- Culotta, A., Kulp, D., & McCallum, A. (2005). Gene Prediction with Conditional Random Fields.

<http://ciir.cs.umass.edu/pubfiles/ir-419.pdf> (accessed 12 May 2021).

Chomsky N (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2, 113-124.

Chomsky N (1959). On certain formal properties of grammars. *Information and Control*, 2, 137-167. https://somr.info/lib/Chomsky_1959.pdf

Clocksin, W. F., & Mellish, C. S. (2003). *Programming in Prolog: Using the ISO Standard* (5th edn.). New York: Springer.

Coden, A. R., Pakhomov, S. V., Ando, R. K., Duffy, P. H., & Chute, C. G. (2005). Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, 38(6), 422-430. <https://doi.org/10.1016/j.jbi.2005.02.009>.

Cunningham, H., Maynard, D., Bontcheva, K., et al. (2011). *Developing Language Processing Components with GATE Version 6 (a User Guide)*. University of Sheffield, Department of Computer Science. <https://gate.ac.uk/releases/gate-6.1-build3913-ALL/tao.pdf> (accessed 13 May 2021).

Daelemans, W., Zavrel, J., Gillis, S., & Berck, P. J. (1996). MBT: A Memory-Based Part of Speech Tagger-Generator. In E. Ejerhed, & I. Dagan (Eds.), *Proceedings of the Fourth Workshop on Very Large Corpora*, 14-27. <https://www.aclweb.org/anthology/W96-0102.pdf>

Danenas, P., Skersys, T., & Butleris, R. (2020). Natural language processing-enhanced extraction of SBVR business vocabularies and business rules from UML use case diagrams. *Data & Knowledge Engineering*, 128, 101822. <https://doi.org/10.1016/j.datak.2020.101822>.

Darroch, J., & Ratcliff, D. (1972). Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*, 43(5), 1470-1480. Retrieved May 12, 2021, from <http://www.jstor.org/stable/2240069>

Deléger, L., Namer, F., & Zweigenbaum, P. (2009). Morphosemantic parsing of medical compound words: transferring a French analyzer to English. *International Journal of Medical Informatics*, 78S, S48-S55.

http://fiamm.free.fr/Publications/IJMI2009_DelegerNamerZweigenbaum.pdf

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), 1-38. https://www.ece.iastate.edu/~namrata/EE527_Spring08/Dempster77.pdf

Dos Santos, C., & Zadrozny, B. (2014). Learning Character-level Representations for Part-of-Speech Tagging. *31st International Conference on Machine Learning, (ICML)*, May 2014. <http://proceedings.mlr.press/v32/santos14.pdf>

Elkan, C (2012). Log-Linear Models and Conditional Random Fields.

<https://cseweb.ucsd.edu/~elkan/250Bwinter2012/loglinearCRFs.pdf> (accessed 12 May 2021).

É. Kiss, K. (1995). Discourse Configurational Languages. Introduction. In: Katalin É. Kiss (szerk.) *Discourse Configurational Languages* (Oxford Studies in Comparative Syntax). New York & Oxford: Oxford University Press, 3–27. <http://grammar.ucsd.edu/courses/lign236/readings/Kiss.pdf>

E-magyar.hu digitális nyelvfeldolgozó rendszer (2015). *emMorph kódlista*. https://e-magyar.hu/hu/textmodules/emmorph_codelist (accessed 13 May 2021).

Forsati, R., & Shamsfard, M. (2014). Hybrid PoS-tagging: A cooperation of evolutionary and statistical approaches. *Applied Mathematical Modelling*, 38 (2014), 3193–3211.

<http://dx.doi.org/10.1016/j.apm.2013.11.047>

Giménez, J., & Márquez, L. (2004). SVMTool: A general POS Tagger Generator Based on Support Vector Machines. In: *Proceedings of the Fourth International Conference on Language Resources and*

- Evaluation (LREC'04)*. European Language Resources Association (ELRA), Lisbon, Portugal.
<http://www.lrec-conf.org/proceedings/lrec2004/pdf/597.pdf>
- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. A. (2011). Part-of-speech tagging for twitter: annotation, features, and experiments. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers*, pages 42–47. Portland, Oregon. <https://www.aclweb.org/anthology/P11-2008.pdf>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780. <https://www.bioinf.jku.at/publications/older/2604.pdf>
- Hutchins W. (2004). *The First Public Demonstration of Machine Translation: the Georgetown-IBM System, 7th January 1954*. Available at:
https://www.researchgate.net/publication/253470063_The_first_public_demonstration_of_machine_translation_the_Georgetown-IBM_system_7th_January_1954
- Kleene, S. C. (1956) Representation of events in nerve nets and finite automata. In: Shannon, C., McCarthy J., (eds.) *Automata Studies*. Princeton, NJ: Princeton University Press.
- Klein D (2010). *CS 288: Statistical Natural Language Processing, Spring 2010*.
<https://people.eecs.berkeley.edu/~klein/cs288/sp10/> (accessed 11 May 2021).
- Kernighan B, & Pike R (1984). *The UNIX Programming Environment*. Englewood Cliffs, NJ: Prentice-Hall. <https://scis.uohyd.ac.in/~apcs/itw/UNIXProgrammingEnvironment.pdf>
- Lafferty, J., McCallum, A., & Pereira F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning (ICML '01)*, June 2001, 282-289.
https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers
- Laurence R. R. et al. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77 (2), 257-286.
<https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorial%20on%20hmm%20and%20applications.pdf>
- Levine JR, Mason T, & Brown D. (1992). *Lex & Yacc*. Sebastopol, CA: O'Reilly & Associates, Inc.
<http://index-of.co.uk/Misc/O'Reilly%20Lex%20and%20Yacc.pdf>
- Manning, C. D. (2011) Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?. In: Gelbukh A.F. (eds) *Computational Linguistics and Intelligent Text Processing. (CICL '11)*. Lecture Notes in Computer Science, vol 6608. Springer, Berlin, Heidelberg.
<https://nlp.stanford.edu/pubs/CICLing2011-manning-tagging.pdf>
- Manning, C. D., Raghavan, P., & Schuetze, H. (2009). *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press. <https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>
- McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML'00)*. June 2000, 591-598. <http://www.ai.mit.edu/courses/6.891-nlp/READINGS/maxent.pdf>
- Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209 (441–458), 415–446.
<https://doi.org/10.1098/rsta.1909.0016>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in

vector space, in: ICLR 2013. <https://arxiv.org/pdf/1301.3781.pdf>

Nadeau, D., Turney, P.D. (2005). A Supervised Learning Approach to Acronym Identification. In: Kégl B., Lapalme G. (eds) *Advances in Artificial Intelligence*. Canadian AI 2005. Lecture Notes in Computer Science, vol 3501. Springer, Berlin, Heidelberg. <https://core.ac.uk/download/pdf/86341.pdf>

Ng, A. Y., & Jordan, M. I. (2002). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, 2. <https://proceedings.neurips.cc/paper/2001/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf>

Nyelvtudomány Kutatóközpont (2014). *Hungarian National Corpus*. http://corpus.nyttud.hu/mnsz/sugo_eng.html#altern (accessed 13 May 2021).

OMG (2017). *Semantics of Business Vocabulary and Business Rules (SBVR) V.1.4*. OMG Doc. No.: formal/2017-05-05. <https://www.omg.org/spec/SBVR/1.4/PDF>

Plank, B., Søgaard, A., & Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, 412-418. <https://www.aclweb.org/anthology/P16-2067.pdf>

Platt, J. C., Cristianini, N., & Shawe-Taylor, J. (1999). Large margin DAGs for multiclass classification. In *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99)*. MIT Press, Cambridge, MA, USA, 547-553.

Pota, M., Marulli, F., Esposito, M., De Pietro, G., & Fujita, H. (2019). Multilingual POS tagging by a composite deep architecture based on character-level features and on-the-fly enriched Word Embeddings. *Knowledge-Based Systems*, 164, 309-323. <https://doi.org/10.1016/j.knosys.2018.11.003>.

Prizadarshi, A., & Saha, S. K. (2020). Towards the first Maithili part of speech tagger: Resource creation and system development. *Computer Speech & Languages*, 62, 101054. <https://doi.org/10.1016/j.csl.2019.101054>

Ptaszynski, M., & Momouchi, Y. (2012). Part-of-speech tagger for Ainu language based on higher order Hidden Markov Model. *Expert Systems with Applications*, 39 (2012), 11576-11582. <http://dx.doi.org/10.1016/j.eswa.2012.04.031>

Ratnaparkhi, A. (1998). Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD Thesis, University of Pennsylvania. <http://www.ai.mit.edu/courses/6.891-nlp/READINGS/adwait.pdf>

Rosen, A. (2006). The annotation system of HunMorph. http://utkl.ff.cuni.cz/~rosen/public/kr_for_ldc.pdf (accessed 13 May 2021).

Sutton, C., & McCallum, A. (2004). *An Introduction to Conditional Random Fields for Relational Learning*. Amherst: University of Massachusetts. <https://people.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf> (accessed 12 May 2021).

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13 (EMNLP '00)*. Association for Computational Linguistics, USA, 63-70. <https://doi.org/10.3115/1117794.1117802>

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL'03)*. Association for Computational Linguistics, USA, 173-180. <https://doi.org/10.3115/1073445.1073478>

Viterbi, A. J. (2009). Viterbi algorithm. *Scholarpedia*, 4(1), 6246. <http://dx.doi.org/10.4249/scholarpedia.6246>

Welch, L. (2003). Hidden Markov Models and the Baum–Welch Algorithm, *IEEE Information Theory Society Newsletter*, 53 (4) (December 2003). <http://www.land.ufrj.br/~sadoc/machinelearning/Baum-Welch.pdf>

Weston, J., and C. Watkins. Support Vector Machines for Multi-Class Pattern Recognition. In *ESANN 1999: Proceedings of the 7th European Symposium on Artificial Neural Networks*, Bruges, Belgium, 21–23 (April 1999), 219–224. <https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es1999-461.pdf>.