Jazykové problémy právních informačních systémů a jejich řešení

Adam Ptašnik

Právnická fakulta (Ústav práva a technologií), MU Brno, Česká republika

Abstrakt v českém jazyce
První část referátu se zabývá jazykovými problémy jako takovými. Nejdříve uvažuje příčiny problémů, a poté ty nejčastější pojmenovává a definuje. Ve druhé části se zabývá různými způsoby jejich řešení. V poslední, třetí, části ukazuje, jak jsou tyto způsoby v informačních systémech zapracovávány, a podle toho je rozděluje do 3 generací řešení. V závěru tyto generace srovnává a z historického vývoje dovozuje vývoj budoucí.

Klíčová slova v českém jazyce
právní informační systém, jazyk, morfologie, synonymie, homonymie

1. Úvod

Právní informační systémy jsou nástrojem, který uživateli pomáhá nalézt řešení jeho právních problémů tak, že dle zadaného dotazu systém vrátí odpověď, která obsahuje právní informace vedoucí k řešení právního problému. Dotaz musí být správně formulován, aby uživatel obdržel odpověď, která skutečně vede k řešení, tedy takovou, kterou očekává. Uživatel, stejně jako autor právních textů uložených v databázi právního informačního systému, ke komunikaci používá přirozený jazyk. Ten je však velmi nedokonalý, neboť není tvořen podle žádných pravidel.Pravidla jeho stavby jsou totiž deskriptivní, nikoli preskriptivní.[1] Úspěch dotazu tedy závisí také na tom, jak je formulován nejen dotaz, ale také jak jsou formulovány právní texty v databázi systému, a jestli formulace dotazu bude formulaci textu odpovídat. Navíc může nastat další komplikace spočívající v tom, že počítač, a tedy i samotný právní informační systém, komunikují v jazyce umělém, vystavěném na přesných pravidlech, a tedy převod přirozeného jazyka do umělého a naopak nemůže být nikdy dokonalý. V tomto článku se seznámíme s některými nejčastějšími problémy plynoucími z přirozeného jazyka, a to morfologií (tvarosloví), synonymií (ekvivalence) a homonymií (ekvivokace), a jejich řešeními.

2. Morfologie

Problematika morfologie spočívá v tom, že mnoho jazyků utváří různé tvary slov dle jejich gramatické funkce.[2] Slovo tak vypadá různě, jestliže se jedná o různý pád, rod nebo např. číslo. Rozdíly obvykle nejsou příliš velké a uživatelé přirozeného jazyka s tímto problém obvykle nemívají, ale stroj, který umí používat pouze jazyk umělý slovo v různých gramatických tvarech zásadně považuje za různé slovo. Při hledání odpovědi v právním informačním systému nám jde především o sémantický význam slov, nikoli o jejich gramatickou funkci. Je nám tedy jedno v jakém tvaru se bude v textu vyskytovat. Pokud tento problém v systému není řešen, systém nenajde všechny výsledky, které hledám.

Možným řešením je tzv. rozšíření slova.[3] Jde o to, že jednotlivé tvary slov se obvykle navzájem příliš neliší, ve velké množství jazyků pravidelně mívají pouze rozdílnou koncovku. V tom případě postačí do vyhledávání zapsat pouze slovo bez koncovky a zadat systému, že má vyhledávat všechna slova začínající na zadaná písmena (pravostranné rozšíření). Pokud se ale jiné tvary slov tvoří i změnou na jiném místě ve slově, pak je třeba písmena, která podléhají změně, nahradit znakem „?“, čímž informujeme počítač, že na tomto místě může být jakékoli písmeno, nebo „*“, kdy zde může být i jakýkoli počet písmen. V tomto případě řešení se však nevyhneme tomu, že systém takto nevytvoří zcela jiné slovo, zvláště u slov krátkých.

Aby uživatel nebyl nucen přemýšlet, na jakých místech může docházet ke změně písmen, byla zavedena technologie fuzzy vyhledávání (obměna slova). Systém se tak pokouší najít nejen zadané slovo, ale také slova obdobná, které se od zadaného liší pouze tím, že má různé jedno písmeno, má jedno písmeno navíc, nebo je některé písmeno vypuštěno. Jelikož se různé tvary vytváří záměnami i několika písmen, obvykle lze nastavit míru obdobnosti vyhledávaných slov, tedy kolik písmen se bude nahrazovat, vypouštět, nebo přidávat.

Změny tvarů slova v jejich různých gramatických funkcích jsou popsány gramatickými pravidly jazyka. Tato pravidla jsou využit u další metody – derivace.[4] Uživatel zadává do vyhledávání slovo v základním tvaru, systém z něj odvodí veškeré morfologické tvary podle derivačních pravidel a ty vyhledává. Derivační pravidla vyplývají z pravidel gramatických, ale nejsou totožná. Gramatická pravidla jsou totiž tvořena pro uživatele přirozeného jazyka, jsou tedy také v jazyce přirozeném. Naopak derivační pravidla musí být v jazyce umělém a musí podchytit veškeré druhy pravidelné tvorby tvarů.[5] Např. v angličtině se tvoří množné číslo přidáním písmene „s“. Tomuto gramatickému pravidlu odpovídají derivační pravidla:

„ ~ => ~s, ~y => ~ies, ~s => ~ses, ~z => ~zes, ...“

Pomocí derivačních pravidel může být vytvořeno mnoho slovních tvarů, které pro dané slovo neexistují, ale se všemi se provádí vyhledávání. Sice zde již téměř neexistuje problém nalezení zcela jiného slova, ale oproti vyhledávání pouze jednoho tvaru se náročnost násobí počtem derivačních pravidel.

Na gramatických pravidlech je založena také lemmatizace.[6] Lemmatizační pravidla jsou zjednodušeně převrácená pravidla derivační. Pomocí nich

1 Ptašnik, A.: Automatizované zpracování právních textů, Ostrava: KEY Publishing, 2007, s. 134

2 Ptašnik, A.: Automatizované zpracování právních textů, Ostrava: KEY Publishing, 2007, s. 27

3 Strossa, P. Vybrané kapitoly z počítačového zpracování přirozeného jazyka. Opava : Slezská univerzita, 1999, s. 27

4 Tamtéž, s. 32

5 Tamtéž, s. 35

6 Tamtéž, s. 39

lze z jakéhokoli tvaru slova vytvořit jeho základní tvar (lemma = základní tvar). Veškeré slova z textů v databázi systému jsou tedy pro účely vyhledávání uloženy v základním tvaru. Slovo zadané do vyhledávání se před vyhledávání do něj převede také. Náročnost je tedy obdobná jako u vyhledávání slova pouze v jednom zadaném tvaru, přitom uživatel může slovo zapsat jakkoli.

3. Synonymie

Synonyma jsou slova, která mají stejný význam, ačkoli jsou různě zapsána. Pokud tento problém není řešen, systém nenajde všechny relevantní výsledky.[7] Obecným řešením je metoda deskriptorů.[8] Slova se rozdělí do skupin tak, že v každé skupině budou ta se stejným významem. Mezi nimi se zvolí jeden zástupce, deskriptor. Pro účely vyhledávání se místo slov z textů databáze uloží jejich deskriptory. Obdobně se pak před vyhledáváním nahrazují zadaná slova pro vyhledávání.

Obecná synonymie není příliš častým jevem.[9] Ačkoli se právníkovi může zdát, že slova mají stejný význam, např. pro básníka bude jejich význam značně odlišný (např. u citově zabarvených slov). Deskriptory je tedy třeba volit vždy dle účelu požití. Existuje však i celá řada specifických synonymií se specifickým řešením.

Synonymii mezi jazyky vyplývající z potřeby vyhledávat odpovědi v několika jazycích najednou (vícejazykové systémy) lze řešit několikanásobným vyhledáváním slova v jednotlivých jazycích po překladu.

Slovo nesprávně zapsané má pro nás stejný význam jako slovo zapsané správně. Jedná se tedy o synonymii. Navíc nesprávně zapsané slovo bývá velmi podobné slovu zapsanému správně, avšak jejich rozdíl je náhodný, nikoli pravidelný. Překlepy, ať již v zadání dotazu nebo v textu v systému lze tedy řešit metodou fuzzy popsanou výše.[10]

Obvykle také nezáleží na velikosti písmen ve vyhledávaných slovech: slovo má stejný význam, jestliže je na začátku věty neb uprostřed, stejně tak, jestliže je zvýrazněno (např. v nadpisu) tak, že je celé napsáno velkými písmeny. Tato synonymie bývá obvykle řešena lemmatizací, kde základním tvarem je slovo zapsané malými písmeny, nebo postupným vyhledáváním možností „všechna velká", „všechna malá" a „první velké" (popř. všech variací velkých a malých písmen ve slově, což je však náročné zejména u delších slov).

Občas se také stejně jako velikost písmen nerozlišuje diakritika. Synonymní jsou pak výrazy s diakritikou a bez ní. Tuto synonymii lze řešit obdobně jako předchozí: lemmatizací, kde lemmatem bude slovo bez diakritiky, nebo postupným vyhledáváním všech kombinací výrazů, kde jsou postupně zaměněna písmena s diakritikou za písmena bez ní a naopak.

Také morfologie je vlastně specifickou synonymií. Pro účely vyhledávání v právních informačních systémech jsou různé tvary stejného slova synonymní. Tato synonymie má svá specifická řešení, o nichž jsme pojednali již výše.

4. Homonymie

Homonyma jsou slova, která jsou zapsána naprosto stejně, ale jejich význam je různý.[11] V případě neřešení tohoto problému tak systém nachází i jiné výsledky, které uživatel nehledal. Základním řešením tohoto problému je kooperace s uživatelem. Pokud systém objeví slovo, které by mohlo mít více významů, položí uživateli doplňující dotaz, které ze slov měl na mysli, a podle odpovědi provádí dále vyhledávání.

Pokud mají shodný zápis jiné gramatické tvary různých slov (homonymie slovních tvarů),[12] pak pro rozhodnutí, o které slovo se jedná lze provést syntaktickou analýzu věty, ve které se vyskytuje. Dle ostatních použitých slov a jejich gramatických tvarů lze dovodit, jakou gramatickou funkci by mělo hledané slovo plnit, a jaký je tedy podle toho očekáván gramatický tvar. Jedná se pak o to slovo, které v tomto tvaru splňuje očekávanou gramatickou funkci.

Ostatní druhy homonymie slov, vč. polysémie (dvě různá slova, která se zapisují naprosto shodně ve všech svých tvarech), lze řešit pomocí kontextové analýzy.[13] Systém prozkoumá slova, která se vyskytují v okolním textu a podle toho rozhodne, se kterým z možných významů je blízký význam okolních slov (podle toho, která slova se obvykle vyskytují společně). Pokud uživatel zadává do vyhledávacího dotazu více slov spojených konjunkcí (nebo obdobnou funkcí), v podstatě tím řeší homonymii touto metodou, neboť tato slova si navzájem určují kontext. Vyhledávají se totiž jen ty texty, kde se tato slova vyskytují společně, tedy v určitém kontextu.

Poslední možností řešení homonymie je metoda analýzy statistické. K jejímu použití potřebujeme statistické údaje o počtu výskytů stejně znějících slov v jejich různých významech. Z této analýzy lze dovodit pravděpodobnost jednotlivých významů. Nelze však úplně vyloučit významy s nižší pravděpodobností, neboť by to uživateli hledajícímu právě tento význam nijak neprospělo. Lze však doporučit toto kriterium zohlednit při řazení výsledků vyhledávání.

5. Vývoj metod řešení vyhledávacích problémů

Metody řešení vyhledávání se v čase vyvíjí, takže právní informační systémy uživateli nabízí stále různé funkce, které může použít, aby jeho práce vedla co nejlépe k výsledku. Aby byl uživatel při vyhledávání úspěšný, musí vědět, co hledá. Musí tedy znát odpověď na některé z otázek:

7 Ptašnik, A.: Automatizované zpracování právních textů, Ostrava: KEY Publishing, 2007, s. 30

8 Strossa, P. Vybrané kapitoly z počítačového zpracování přirozeného jazyka. Opava : Slezská univerzita, 1999, s. 68

9 srov. pojem *silná synonymie* v Ptašnik, A.: Automatizované zpracování právních textů, Ostrava: KEY Publishing, 2007, s. 30

10 Strossa, P. Vybrané kapitoly z počítačového zpracování přirozeného jazyka. Opava : Slezská univerzita, 1999, s. 225

11 Ptašnik, A.: Automatizované zpracování právních textů, Ostrava: KEY Publishing, 2007, s. 28

12 srov. *morfologická homonymie* v Strossa, P. Vybrané kapitoly z počítačového zpracování přirozeného jazyka. Opava : Slezská univerzita, 1999, s. 88

13 Strossa, P. Vybrané kapitoly z počítačového zpracování přirozeného jazyka. Opava : Slezská univerzita, 1999, s. 88 a n.

Co je správná odpověď? Při zobrazení výsledku vyhledávání musí poznat, zda se jedná o odpověď, která vede k řešení jeho problému, či nikoli.

Jakou metodou systém vyhledává? Aby mohl uživatel správně vyhledávání nastavit a správně zadat dotaz pro vyhledávání, tj. tak, aby vyhledávání bylo co nejefektivnější, musí vědět, jaká je použitá vyhledávací metoda.

Jak přesně vypadá správná odpověď? Uživatel musí dokonce správnou odpověď přesně znát, aby mohl přesně takový dotaz, který vede k této odpovědi.

Dle nutnosti znát odpovědi na předložené otázky si můžeme s jistou mírou nepřesnosti výše uvedené vyhledávací metody rozdělit do několika generací.

5.1 Nultá generace metod, tj. neřešení výše uvedených problémů.

Uživatel zde musí správnou odpověď nejen poznat, ale musí ji předem znát. Musí např. vědět, v jakém tvaru je v odpovědi použito vyhledávané slovo, neboť nemá žádnou možnost tento problém řešit. V těchto případech se používá uživatelská metoda „střelby", kdy sám uživatel zkouší různé možnosti, jak by odpověď mohla vypadat.

Vyhledávací formulář je v této generaci velmi jednoduchý, neboť obsahuje pouze pole pro zadání dotazu a tlačítko k jeho odeslání.

5.2 První generace

V tomto případě uživatel kromě toho, že pozná, zda se jedná o správnou odpověď, navíc musí poměrně podrobně znát používanou metodu vyhledávání, aby mohl každý jednotlivý dotaz konstruovat tak, aby byla tato funkce správně použita. Příkladem jsou metody rozšíření, kooperace s uživatelem a kontextová analýza, kdy uživatel musí vědět, kam a jaký zástupný znak má do dotazu umístit, jaký je význam zadávaného slova, popř. jaké ještě slovo zadat, aby to původní bylo vyhledáno ve správném významu.

Ačkoli je zde třeba zadávat např. zástupné znaky, v tomto směru se formulář příliš nezmění. Pouze se zvýší počet polí pro zadávání dotazu tak, aby bylo možné každou část dotazu vyhledávat pomocí jiné metody, přidají přepínače pro zapínání těchto metod a pole pro výběr systémem nabízených možností při kooperaci s uživatelem.

5.3 Druhá generace

Při použití těchto metod již uživateli stačí, že pozná, zda se jedná o správnou odpověď. Musí o ní mít tedy určité povědomí, podle kterého dotaz zadává. Do této generace lze zařadit metody lemmatizace, derivace, různé obměny slov jako je fuzzy, překlady do jiných jazyků a syntaktickou analýzu.

Oproti formuláři první generace se přidají pouze další přepínače na zapínání jednotlivých metod vyhledávání a různé volby nastavení těchto funkcí, které nemusí být nutně přímo ve vyhledávacím formuláři, protože se obvykle nastavují pro uživatele, nikoli pro každý jednotlivý dotaz (např. míra obdobnosti slov při fuzzy vyhledávání).

5.4 Třetí generace

Poslední známá generace vyhledávacích funkcí při jejich dokonalé funkčnosti odstraňuje poslední nutnou znalost uživatele. Ten tedy ani nemusí poznat, zda odpověď, které se mu dostane, je správná (tedy taková, jakou hledal). Do této generace lze zařadit metodu relevance, tedy odstraňování nesmyslných variant (používá se např. při lemmatizaci v podobě stop slov), a statistické analýzy, tesy jakési míry relevance. Uživatel tedy získává odpověď, která je podle zadaného dotazu pravděpodobně nejdůležitější, popř. několik odpovědí seřazených podle důležitosti. Záleží pouze na konkrétní implementaci této metody, jaká kriteria bere systém v úvahu, když přiděluje výsledkům míru relevance. Systémy, které s těmito metodami pracují, by měly tato kriteria volit v závislosti na jejich určení, neboť relevance je relativní.

Vyhledávací formulář se skládá pouze z pole pro zadání dotazu a tlačítka k jeho odeslání. Ostatní již zajistí samotná metoda bez zásahu uživatele.

6. ZÁVĚR

Z přehledu generací funkcí lze usoudit, že právní informační systémy vyhledávají stále efektivněji, odpovědi na dotazy uživatelů jsou stále více použitelné pro řešení jejich právních problémů. Z pohledu práce uživatele s právním informačním systémem lze vývoj rozdělit na 2 etapy, kdy v první přibývaly funkce, které mohl uživatel dle svého rozhodnutí nastavit a použít. Byl ale nucen je správně pochopit a zvládnout, aby je mohl využívat efektivně. Ve druhé etapě se již od těchto funkcí ustupuje a směřuje se k metodám, které uživatele nijak nezatěžují (vlastně ani nepozná, že jsou použity), nemusí (ale často ani nemůže) znát, jak fungují, přitom vyhledávání zefektivňují. Zatímco v první etapě stály v popředí zájmu tvůrců vědy jako matematika, statistika a lingvistika a dbalo se především na to, aby metodu zvládl počítač, ve druhé etapě již do popředí vystupují vědy jako psychologie, sociologie a právo a dbá se na to, aby použitá metoda nezatěžovala uživatele. Tvůrce systému tak musí zkoumat, co a jakým způsobem bude jeho budoucí uživatel vyžadovat, již při jeho vývoji a ne až při zadání dotazu.

Literatura:

Ptašnik, A.: Automatizované zpracování právních textů, Ostrava: KEY Publishing, 2007, 167 stran, ISBN: 978-80-87071-47-2

Strossa, P. Vybrané kapitoly z počítačového zpracování přirozeného jazyka. Opava : Slezská univerzita, 1999, 277 stran, ISBN: 80-7248-041-3