DOI: 10.15503/onis2015.212.220

Zastosowanie gier skierowanych na cel do anotacji korpusów językowych

Dagmara Dziedzic, dag.dziedzic@gmial.com
Uniwersytet im. Adama Mickiewicza w Poznaniu
Ul. Wieniawskiego 1, 61-712 Poznań

Wojciech Włodarczyk, wlodarczyk.woj@gmial.com
Uniwersytet im. Adama Mickiewicza w Poznaniu
Ul. Wieniawskiego 1, 61-712 Poznań

Streszczenie

Istnienie problemów AI-zupełnych przyczyniło się do poszukiwań alternatywnych sposobów rozwiązywania problemów sztucznej inteligencji, nie opartych wyłącznie na pracy komputera. Pomimo że komunikacja jest dla ludzi czymś oczywistym, nadal nie istnieje sposób jej automatyzacji. Aktualnie powszechnie stosowanym podejściem w rozwiązywaniu problemów NLP jest podejście statystyczne, którego powodzenie zależy od wielkości korpusu językowego. Przygotowanie rzetelnego zbioru danych jest zatem kluczowym aspektem tworzenia statystycznego systemu sztucznej inteligencji. Z uwagi na zaangażowanie specjalistów jest to proces czasochłonny i kosztowny. Jednym z obiecujących podejść, pomagających zredukować czas i koszt tworzenia otagowanego korpusu, jest korzystanie z gier skierowanych na cel. Ambicją niniejszej pracy jest przybliżenie poszczególnych etapów tworzenia gry przeznaczonej do pozyskania zasobów językowych oraz omówienie skuteczności jej działania. Analiza ta zostanie przeprowadzona na podstawie kolekcji gier Wordrobe wspierających anotacje korpusu języka naturalnego.

Słowa kluczowe: gry skierowane na cel, GWAP, crowdsourcing, human computation, przetwarzanie języka naturalnego, sztuczna inteligencja, AI-zupełne, anotacja korpusu, Wordrobe

Applications of games with a purpose to obtain annotated language resources

Abstract

The existence of AI-complete problems has led to growth in research of alternative ways for solving artificial intelligence problems, which could not be solved on the computer. Although communication is obvious for people, there is still no way for its automation. Statistical approach became widely used in solving the problems of NLP. One of the main factors of its success is the size of the training corpus. Preparing a reliable set of data is therefore a key aspect of creating an artificial intelligence statistical system. Due to the involvement of a large number of specialists it is a very time-consuming and expensive process. One of the promising approaches to help reduce the time and cost of creating a tagged corpus is the use of games with a purpose. The objective of the following paper is to present the stages of creating games with a purpose used for obtaining annotated language resources and to discuss its effectiveness. This analysis will be done based on project Wordrobe, the collection of games created to support gathering annotated corpus of natural language.

Key Words: game with a purpose, GWAP, crowdsourcing, human computation, natural language processing, artificial intelligence, AI-complete, corpus annotation, Wordrobe

Podstawowa umiejętność liczenia towarzyszyła ludziom niemal od zawsze. Próby wyręczenia człowieka w tej czynności lub automatyzacji prowadzonych przez niego obliczeń sięgają co najmniej dziejów starożytnych. Wówczas powstawały pierwsze urządzenia liczące tj. liczydła, które z czasem zostały zastąpione przez kalkulatory, a następnie z biegiem wieków, przez współczesne komputery[1].

Do tej pory uważano, że brak możliwości usprawnienia procesu rozwiązywania problemów obliczeniowych jest bezpośrednio związany z niedoskonałością stworzonych przez człowieka maszyn. Jednakże, pomimo wciąż zwiększającej

1 M. M. Sysło, *Historia rachowania – ludzie, idee, maszyny. Historia mechanicznych kalkulatorów*, [w:] Tenże (red.), *Homo Informaticus*, Warszawa 2012, s. 267-271.

się mocy obliczeniowej komputerów, nadal istnieje wiele problemów, które będąc trywialne dla ludzi, nie mogą być rozwiązane przy użyciu komputera. Na gruncie sztucznej inteligencji, problemy te są określane mianem problemów AI-zupełnych[2] (ang. *AI-complete*)[3]. Nie są one redukowalne do problemu nie będącego innym problemem sztucznej inteligencji. Nie tylko rozwiązanie problemu AI-zupełnego jest skomplikowanym procesem, ale również ocena samego rozwiązania jest problematyczna. Wymaga ona interpretacji wykonanych obliczeń, która związana jest z posługiwaniem się wiedzą zdroworozsądkową.

Przykładowo, mowa będąca podstawowym środkiem komunikacji, nie jest związana tylko z posiadaniem wiedzy z zakresu lingwistyki. Obejmuje ona również rozmówców i wszystkie cechy, które oni posiadają oraz środowisko komunikacyjne, w którym przekazywana jest informacja[4]. Pomimo tak złożonej formy, komunikacja jest czymś naturalnym i bezproblemowym dla człowieka. Ludzie z łatwością dekodują tak przekazywaną informację, jednakże proces automatyzacji tego zachowania nie jest wciąż dobrze określony.

Warto zauważyć, że samo zdekodowanie informacji nie pociąga za sobą rozumienia treści, którą ona ze sobą niesie. Przetwarzanie języka naturalnego (ang. *natural language processing*, NLP) obejmuje nie tylko proces komunikowania się, ale również kontekst, w którym są osadzone słowa[5]. Problem stanowią często niejednoznaczności pojawiające się w języku, w przypadku których nie wystarczy rozumienie sensu pojedynczych słów, ale dopiero przy uwzględnieniu szerokiego kontekstu, wypowiedź nabiera właściwego znaczenia[6]. Z powyższych powodów, przetwarzanie języka mówionego i rozumienie języka naturalnego, jak również inne aspekty komunikacji międzyludzkiej, zalicza się do grupy problemów AI-zupełnych.

Podejścia do rozwiązywania problemów przetwarzania języka naturalnego

Teoretyczne podstawy NLP wywodzą się z lingwistyki komputerowej, dziedziny nauki zajmującej się testowaniem hipotez dotyczących mowy i języka przy użyciu modeli komputerowych[7]. Do lat osiemdziesiątych XX w. istniało przekonanie mówiące, że skoro ludzie są w stanie opisać język przy pomocy reguł gramatycznych, to wystarczy użyć tych reguł do odtworzenia umiejętności posługiwania się językiem w postaci sztucznej inteligencji[8]. Początkowo próby stworzenia takiej sztucznej inteligencji opierały się na np. poszukiwaniu uniwersalnej gramatyki, która za pomocą reguł opisywałaby szyk zdania, rolę i znaczenie słów[9]. W praktyce okazało się, że tworzenie reguł opisujących wszystkie przypadki użycia języka nie jest możliwe. W przeciwieństwie do języków formalnych, konstruowanie zdań w języku naturalnym charakteryzuje się dużą swobodą. Dodatkowo, w języku naturalnym występuje duża liczba wyjątków i niejednoznaczności. Przede wszystkim jednak brak w nim ścisłych reguł gramatycznych, które mogłyby umożliwić jego formalizację. Próby bardziej sformalizowanego opisu języka generowały bardzo dużą liczbę reguł, które często kolidowały ze sobą.

Z uwagi na niepowodzenia związane z próbami odtworzenia języka przy pomocy reguł i możliwości, jakie zaczęły nam dostarczać komputery, zaczęto używać statystycznych metod przetwarzania języka naturalnego[10]. Podejście statystyczne (ang. *statistical-based approach*), które opiera się na dużej liczbie danych i automatycznym wyszukiwaniu relacji pomiędzy nimi, jest aktualnie powszechne stosowane w dziedzinach związanych z NLP[11]. Podejście to polega na stworzeniu modelu opisującego zasady językowe w oparciu o duży zbiór tekstów, który w lingwistyce nazywany jest korpusem. Metody statystyczne często bazują na uczeniu maszynowym, podczas którego, w oparciu o dużą liczbę poprawnych

2 Z uwagi na brak jednoznacznego tłumaczenia terminu w języku polskim, w niniejszym artykule stosowane będzie tłumaczenie wprowadzone analogiczne do polskiego określenia „problemy NP-zupełne", używanego w kategoriach teorii obliczeń. Zob. J. Grytczuk, *Czy wszystko można obliczyć. Łagodne wprowadzenie do złożoności obliczeniowej*, [w:] M. M. Sysło (red.), *Homo Informaticus*, Warszawa 2012, s. 109-110.

3 R. V. Yampolskiy, *AI-Complete, AI-Hard, or AI-Easy – Classification of Problems in AI*, [w:] S. Visa, A. Inoue, A. Ralescu (red.), *Proceedings of the 23rd Midwest Artificial Intelligence and Cognitive Science Conference*, Cincinnati 2012, s. 94.

4 A. G. Adami, *Automatic Speech Recognition: from the Beginning to Portuguese Language*, Materiały konferencyjne, International Conference on Computational Processing of Portuguese 2010, https://www.inf.pucrs.br/~propor2010/proceedings/tutorials/Adami.pdf, 29.12.2014, s. 1.

5 J. Szymański, W. Duch, *Annotating Words Using WordNet Semantic Glosses*, [w:] T. Huang, Z. Zeng, C. Li, C. S. Leung (red.), *Neural Information Processing*, Doha 2012, s. 180-181.

6 Tamże, s. 186-187.

7 I. A. Bolshakov, A. Gelbukh, *Computational Linguistics: Models, Resources, Applications*, Mexico 2004, s. 25-26.

8 K. S. Jones, *Natural Language Processing: a Historical Review*, „Artificial Intelligence Review" 2001, s. 6-8.

9 Tamże, s. 3.

10 Popularne jest również stosowanie podejścia hybrydowego (ang. *hybrid approach*), w którym wykorzystuje się zarówno metody symboliczne (w tym regułowe) jak i statystyczne. Zob. P. Resnik, J. Klavans, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, New Mexico 1994.

11 P. M. Nadkarni, L. Ohno-Machado, W. W. Chapman, *Natural Language Processing: an Introduction*, „The Journal of the American Medical Informatics Association" 2011, no. 18 (5), s. 544-545.

przykładów dostarczonych w postaci korpusu, sztuczna inteligencja uczy się wykonywać dane zadanie[12]. Metody statystyczne nie są zależne od języka i mogą być z powodzeniem stosowane do większości popularnych problemów NLP[13].

Złoty standard anotacji

Wypracowanie rozwiązania problemów sztucznej inteligencji, w tym problemów AI-zupełnych, skupia się nie tylko na znalezieniu optymalnych metod statystycznych, ale również na zebraniu wysokiej jakości korpusu językowego. Zbiór tekstów, który wykorzystywany jest w procesie uczenia maszynowego do trenowania i ewaluacji algorytmów sztucznej inteligencji nazywany jest złotym standardem (ang. *gold standard*)[14]. Przygotowanie takiego korpusu polega na ręcznej anotacji danych wykonywanej przez specjalistów. Złoty standard jest traktowany jako wzorzec, z którym porównywane są efekty pracy sztucznej inteligencji, w celu ewaluacji jej działania[15]. Proces tworzenia takiego korpusu jest czasochłonny i kosztowny. Stworzenie wysokiej jakości korpusu jest kluczowe dla rozwiązywania problemów NLP, dlatego często jest on niezależnie anotowany przez kliku specjalistów. Efekty ich pracy są następnie porównywane, a do złotego standardu wybierane są te zdania, które mają największą zgodność pomiędzy ekspertami. Zapotrzebowanie na dużą liczbę odpowiednio otagowanych danych jest głównym minusem podejścia statystycznego. Aby usprawnić proces i obniżyć koszty tworzenia złotego standardu poszukuje się alternatywnych metod pozyskiwania wysokiej jakości zasobów lingwistycznych.

Alternatywne metody pozyskiwania zasobów lingwistycznych

Zazwyczaj, kiedy ludzie stoją przed zadaniem wykonania skomplikowanych obliczeń sięgają po pomoc komputera. Po dostarczeniu sformalizowanego opisu problemu i zastosowaniu odpowiedniego programu, otrzymują w krótkim czasie rozwiązanie. W przypadku rozwiązywania problemów AI-zupełnych, role te odwracają się. Komputer potrzebuje rozwiązania, które zazwyczaj opracowane zostaje przez dużą grupę ludzi. Korzystanie z mądrości tłumu daje lepsze rezultaty, niż gdyby dany problem rozwiązywały jednostki. Duża i zróżnicowana społeczność znajduje kreatywniejsze rozwiązania lub tworzy innowacyjne koncepcje, dzięki lepszej ocenie rzeczywistości[16]. Samo rozwiązanie natomiast nie jest efektem osiągnięcia konsensusu w grupie, lecz rezultatem gromadzenia dużej liczby indywidualnych pomysłów[17]. Warte podkreślenia jest to, że omawiana grupa ludzi nie składa się w większości z ekspertów. Pomimo tego, szczególnie w przypadku prostych zadań, jakość wypracowanego w ten sposób rozwiązania jest porównywalna z jakością, którą osiągnęliby specjaliści.

Pomysł zaangażowania dużej grupy ludzi do wykonywania niewielkich zadań obliczeniowych nie jest nowy. W połowie XVIII w., przed erą komputerów, takie zespoły rozwiązywały problemy ówczesnych społeczeństw, w podobny sposób, w jaki dziś prowadzi się obliczenia w paradygmacie *human computation* (HC), czy *crowdsourcingu*[18]. Współczesne znaczenie idei HC nadał w 2005 r. Luis von Ahn, definiując ją jako paradygmat wykorzystujący ludzkie zdolności do rozwiązywania problemów, których komputery nie są w stanie jeszcze rozwiązać[19]. HC opisuje sytuację, w której systemy komputerów i duża liczba osób wspólnie pracują nad rozwiązaniem problemu, który nie mógłby być rozwiązany ani przez samych ludzi, ani przez sam komputer. Ta definicja nie obejmuje jednak wszystkich technologii, dzięki którym ludzie współpracują z komputerami. Podobnym do HC i niewiele starszym od niego konstruktem jest pojęcie *crowdsourcingu*, które po raz pierwszy zostało użyte przez Jeffa Howe'a w roku 2006[20]. *Crowdsourcing* jest procesem, w którym zadania wykonywane

12 Tamże, s. 546.

13 Tamże, s. 545.

14 P. Paroubek, S. Chaudiron, L. Hirschman, *Principles of Evaluation in Natural Language Processing*, „French Association for Natural Language Processing" 2007, no. 48(1), s. 5-7.

15 Pojęcie „standard złota" zostało zaczerpnięte z ekonomii i oznacza pierwszy międzynarodowy system pieniężny, w którym każdej walucie przyporządkowana była określona waga złota. Zob. K. Samuel, *The Gold Standard and the Origins of the Modern International Monetary System*, „Centre Études internationales et Mondialisation" 2003, no. 3(1), s. 6-7. Obecnie, w różnych dziedzinach nauki, złotym standardem określa się metody i informacje przyjmowane jako normy.

16 J. Surowiecki, *The Wisdom of Crowds*, New York 2005.

17 Tamże.

18 D. A. Grier, *Foundational Issues in Human Computing and Crowdsourcing*, „Conference on Human Factors in Computing Systems" 2011, s. 7-8.

19 L. von Ahn, *Human Computation*, Rozprawa doktorska, Carnegie Mellon University 2005, http://reports-archive.adm.cs.cmu.edu/anon/2005/CMU-CS-05-193.pdf, 29.12.2014.

20 J. Howe, *The Rise of Crowdsourcing*, „Wired Magazine" 2006, no. 14 (6), s. 1-4.

tradycyjnie przez pracowników są rozsyłane poprzez otwarte zgłoszenie (ang. *open call*) do dużej, niezdefiniowanej grupy ludzi[21]. W praktyce *crowdsourcing* jest metodą rozwiązywania problemów poprzez wykorzystanie Internetu, który pozwala na komunikowanie się z dużą liczbą osób[22]. Obecnie pojęcia HC i *crowdsourcing* są często używane synonimiczne. Jednakże istnieją różnice pomiędzy nimi. Podczas gdy w HC komputery zastępuje się niezdefiniowaną grupą ludzi, w *crowdsourcingu*, zastępuje się nią tradycyjnych pracowników[23]. HC jest środkiem służącym do rozwiązywania problemów obliczeniowych, natomiast takie problemy tylko czasami są rozwiązywane dzięki wykorzystaniu drugiej metody [24]. Obecnie oba podejścia są z powodzeniem używane, jako metody pozyskiwania danych - również, gdy potrzebna jest ich wysoka jakość. W przypadku tworzenia złotego standardu anotacji takimi danymi jest odpowiednio zanotowany korpus językowy. Przy tworzeniu korpusu z wykorzystaniem *crowdsourcingu* można skorzystać z trzech typów rozwiązań:

- siły roboczej (ang. *mechanised labour*),
- altruistycznego *crowdsourcingu*,
- gier skierowanych na cel (ang. *games with a purpose, GWAPs*)[25].

Siłą roboczą są pracownicy, którzy odpowiedzieli na otwarte zgłoszenie. W zamian za rozwiązanie małych zadań otrzymują niewielką zapłatę. Altruistyczny *crowdsourcing* opiera się na dobrej woli wolontariuszy, ponieważ za rozwiązanie zadań nie otrzymują nic w zamian. Natomiast gry skierowane na cel prezentują zadanie w postaci gry. Ideę gier tego typu, rozwijających paradygmat HC, po raz pierwszy przedstawił L. von Ahn tworząc ESP Game[26]. Dwóch graczy było proszonych o opisanie obrazka. Jeżeli oboje użyli do opisu tych samych słów byli nagradzani punktami. Poprawne odpowiedzi zostały zapamiętane a następnie użyte do ulepszenia algorytmu wyszukiwania obrazków w wyszukiwarce internetowej. Gry skierowane na cel znajdują swoje zastosowanie w różnych dziedzinach nauki, w których istnieją problemy niedające się rozwiązać przy użyciu samego komputera. Zyskały one popularność np. w rozwiązywaniu problemów biologicznych. Przy ich pomocy oznacza się neurony znajdujące się w siatkówce oka[27], rozwiązuje się problemy związane ze zwijaniem się RNA[28], czy szuka się wzorców w sekwencjach nukleotydów[29].

Sukces ESP Game i innych gier skierowanych na cel pokazuje, że odpowiednio zachęceni ludzie mogą stać się częścią zbiorowych obliczeń. Podejście polegające na tworzeniu tego typu gier ma ogromny potencjał. Nie wymaga opłacania pracowników, a forma, w jakiej przedstawione są zadania sama zachęca do ich wykonywania. Technika ta bazuje na przyjemności, która płynie z osiągania kolejnych celów, rywalizacji lub współpracy. Monotonna praca ekspertów polegająca na oznaczaniu danych, zostaje poddana procesowi grywalizacji. Grywalizacja polega na użyciu mechanizmów znanych z gier w celu zwiększenia zaangażowania ludzi do wykonywania zadań, nawet jeśli uważane są one za nudne lub rutynowe[30]. Wykorzystując w ten sposób czas, który ludzie spędzają w grach online, można otrzymać dane trenujące, które następnie mogą posłużyć do rozwiązania problemów AI-zupełnych. W przypadku poblemów NLP ich rozwiązanie związane jest z zebraniem oznaczonego metadanymi korpusu językowego. Istotne jest więc efektywne zgrywalizowanie procesu anotacji, co pozwala w pełni wykorzystać potencjał, jaki niesie ze sobą idea HC. Doskonałym przykładem zastosowania gier skierowanych na cel jest projekt Wordrobe składający się z kolekcji gier wspierających anotacje korpusu języka naturalnego[31].

Proces pozyskiwania zasobów językowych przy użyciu *crowdsourcingu*

Projektowanie gier skierowanych na cel jest podobne do projektowania algorytmu. A to dlatego, że gra taka powinna umożliwiać analizowanie zarówno poprawności jak i skuteczności pozyskiwania danych[32]. Zasadniczo proces pozyskiwania zasobów językowych przy użyciu metod *crowdsourcigu* można podzielić na cztery etapy:

---

21 J. Howe, *Crowdsourcing: A Definition*, http://crowdsourcing.typepad.com, 31.12.2014.

22 Przykładem takiego narzędza jest platforma Amazon Mechanical Turk: https://www.mturk.com, 30.12.2014.

23 A J. Quinn, B. B. Bederson, *Human Computation: a Survey and Taxonomy of a Growing Field*, „Conference on Human Factors in Computing Systems" 2011, s. 1405.

24 Tamże, s. 1403.

25 M. Sabou, K. Bontcheva, L. Derczynski, A. Scharl, *Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines*, „Language Resources and Evaluation Conference" 2014, s. 859-860.

26 L. von Ahn, *Games With A Purpose*, "Institute of Electrical and Electronics Engineers Computer Magazine" 2006, s. 96-97.

27 Gra Eyewire, http://eyewire.org ,30.12.2014.

28 Gra Eterna, http://eternagame.org/, 30.12.2014.

29 Gra Phylo, http://phylo.cs.mcgill.ca/, 30.12.2014.

30 P. Tkaczyk, *Grywalizacja. Jak zastosować mechanizmy gier w działaniach marketingowych*, Gliwice 2012, s.10.

31 Zestaw gier dostępny jest pod adresem: www.wordrobe.org, 30.12.2014.

32 L. von Ahn, *Games…*, dz. cyt., s. 96.

1. Definicje projektu (ang. *project definition*)
2. Przygotowanie danych i platformy *crowdsourcingowej* (ang. *data preparation*)
3. Realizacja projektu (ang. *project execution*)
4. Ewaluacja i agregacja wyników (ang. *data evaluation and aggregation*)[33].

Etapy te zostały opracowane na podstawie meta-analizy wielu projektów *crowdsourcingowych* uwzględniając zarówno siłę roboczą, altruistyczny *crowdsourcing* i gry skierowane na cel. Stanowią one swego rodzaju dobre praktyki pomagające zaprojektować dowolny system *crowdsourcingowy*, który ma służyć pozyskiwaniu zasobów językowych. W dalszej części artykułu powyższe etapy[34] zostaną szerzej omówione i umiejscowione w projekcie Wordrobe.

Etap pierwszy: Definicja projektu

Po wyborze odpowiedniego typu *crowdsourcingu*, należy zdefiniować projekt, czyli określić, jakiego typu zasoby będą gromadzone poprzez stworzone narzędzie. W przypadku problemów NLP sprowadza się to do określenia, jakiego typu dane będą oznaczane w korpusie. Zadania, które wykonują użytkownicy powinny zostać zaprojektowane tak, aby mogły być wykonywane przez osoby, które nie są ekspertami. Powinny bazować na intuicjach użytkownika dotyczących posługiwania się językiem, zamiast odwoływać się do wiedzy specjalistycznej. Projektując takie zadania, powinno się unikać pytań otwartych, gdyż zdecydowanie lepsze rezultaty uzyskuje się po ograniczeniu odpowiedzi do kilku możliwości[35]. Wiele problemów NLP można sprowadzić do problemu klasyfikacji, w którym wybór odpowiedniej anotacji ograniczony jest do pewnej, skończonej liczby kategorii. Ostatnim etapem definiowania projektu jest określenie nagrody za wykonanie zadania i sposobu oceny poprawności jego rozwiązania. W zależności od tego, czy uczestnicy otrzymują nagrodę pieniężną, wykonują zadania na podstawie wolontariatu, czy główną formą nagrody są punkty, ich zaangażowanie może być różne, co przejawia się w jakości przesyłanych rozwiązań[36].

Wspomniany wcześniej projekt Wordrobe[37] spełnia powyższe kryteria. Jest to platforma składająca się z siedmiu niewielkich gier, z których każda przeznaczona jest do anotacji jednego typu metadanych. Każda z gier zaprojektowana jest w ten sam sposób i może być używana przez osoby niebędące ekspertami w dziedzinie językoznawstwa. Aktualnie dostępne są gry, w których zadaniem gracza jest:

- anotacja części mowy,
- anotacja znaczenia słowa,
- określenie koreferencji,
- określenie tematu zdania,
- klasyfikacja nazw własnych,
- klasyfikacja przyimków,
- zaklasyfikowanie rzeczowników do grup (żywotny i nieżywotny).

Rozwiązanie każdego z powyższych problemów polega na odpowiedzi na pytanie jednokrotnego wyboru.

W celu zachęcenia gracza do wykonywania anotacji, wykorzystano dwa typy nagród: szuflady (ang. *drawers*)[38] i punkty (ang. *points*). Pierwszą z nich otrzymuje się za wykonanie pewnego zbioru zadań. Im zadania są trudniejsze, tym mniej należy ich wykonać, aby zdobyć nagrodę. Drugim typem nagród są punkty, które otrzymuje się za wykonanie każdego z zadań, natomiast ich ilość zależy od poprawności rozwiązania. Jest ona obliczana na podstawie zgodności użytkownika z odpowiedziami innych graczy i aktualną wersją korpusu. Podczas odpowiadania na pytanie uczestnik gry każdorazowo może oznaczyć stopień, w jakim jest przekonany, co do prawidłowości swojej odpowiedzi. Jeżeli gracz jest pewny swojej odpowiedzi i odpowie prawidłowo otrzymuje więcej punktów za wykonanie zadania. Co więcej zauważono wysoką zgodność pomiędzy stopniem pewności użytkownika a poprawnością udzielanych przez niego odpowiedzi.

33 M. Sabou, K. Bontcheva, L. Derczynski, A. Scharl, dz. cyt., s. 860.

34 W dalszej części artykułu kolejne etapy projektowania narzędzia crowdsourcingowego zostały opracowane na podstawie: M. Sabou, K. Bontcheva, L. Derczynski, A. Scharl, dz. cyt., s. 860-866.

35 A. Aker, M. El-Haj, M. Albakour, U. Kruschwitz, *Assessing Crowdsourcing Quality through Objective Tasks*, „Language Resources and Evaluation Conference" 2012, s. 1460-1461.

36 Tamże, s. 1460-1461.

37 W dalszej części artykułu, opis projektu Wordrobe został opracowany na podstawie: N. J. Venhuizen, V. Basilem, K. Evang, J. Bos, *Gamification for Word Sense Labeling*, „IWCS" 2013, 397-403.

38 Nazwa platformy „Wordrobe" nawiązuje do angielskiego słowa „*wardrobe*", które w języku polskim oznacza dosłownie „szafę" - stąd nagrodami w grze są szuflady.

Etap drugi: Przygotowanie danych i platformy crowdsourcingowej

Drugi etap skupia się na pozyskaniu i przygotowaniu danych do anotacji. W przypadku rozwiązywania problemów NLP, takimi danymi jest korpus lingwistyczny. W zależności od celu pozyskiwania zasobów językowych używa się czystego korpusu, na którym nie oznaczono żadnych metadanych lub korpusu preanotowanego, którego anotacja w procesie *crowdsourcingu* jest poprawiana. Drugie podejście jest używane powszechnie, ponieważ ułatwia ewaluację wyników. Zgodność anotacji użytkowników można wówczas porównywać z aktualną wersją korpusu.

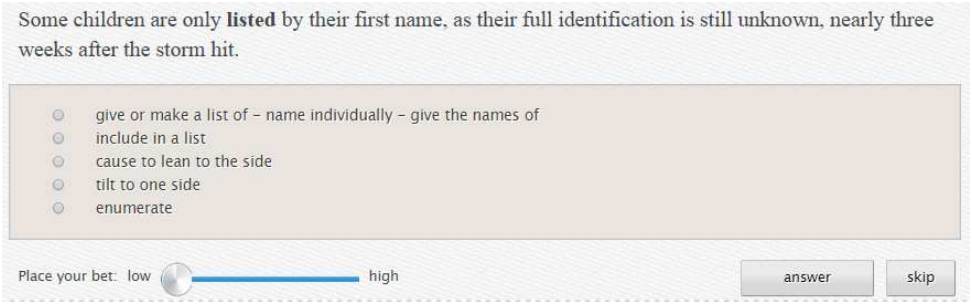
Ze względu na strukturę zadań *crowdsourcingowych*, korpus należy podzielić na pojedyncze zadania, składające się z kilku zdań lub niedługich tekstów. Jeżeli proces pozyskiwania danych nie zakłada użycia gotowego narzędzia[39], należy wówczas przygotować platformę *crowdsourcingową*. Na taką platformę składają się zazwyczaj dwa podsystemy: panel użytkownika i panel zarządzający. Pierwszy z nich to część interfejsu, która dostępna jest tylko dla użytkownika. To w niej następuje anotacja danych, dlatego powinna być przejrzyście zaprojektowana i posiadać przyjazną formę graficzną. Z platformy zazwyczaj nie korzystają specjaliści, dlatego warto, aby zachęcała ona do wykonywania zadań. Z tego powodu należy zadbać nie tylko o czytelność interfejsu, ale również o dostarczenie użytkownikom instrukcji w ich języku natywnym.

W przypadku projektowania gry skierowanej na cel, ważnym elementem, który należy uwzględnić przy jej tworzeniu, jest atrakcyjność rozgrywki. Gra powinna mieć jasno określone zasady i stanowić wyzwanie dla graczy. Aby to osiągnąć należy zadbać o losowość rozgrywki, odpowiednią punktację i rankingi graczy oraz – gdy zaistnieje taka potrzeba- wprowadzić limity czasu wykonania określonego zadania[40].

Drugim, wspomnianym wcześniej, elementem platformy *crowdsourcingowej* jest panel zarządzania zawierający narzędzia służące do monitorowania przebiegu anotacji zasobów językowych. Pozwala on nadzorować działania użytkowników i gromadzić informacje, o jakości ich pracy. Dzięki temu istnieje możliwość ciągłej poprawy procesu *crowdsourcingu*, szczególnie w przypadku iteracyjnego modelu pozyskiwania danych[41]. Iteracyjny model zakłada przeprowadzanie niewielkich, zamkniętych etapów *crowdsourcingu*, po których zakończeniu analizuje się wyniki, obejmujące zarówno pracę użytkowników, jak i napotkane błędy. Po przeprowadzeniu analizy i wprowadzeniu poprawek rozpoczyna się kolejny etap *crowdsourcingu*.

Gry dostępne na platformie Wordrobe opierają się na korpusie *The Groningen Meaning Bank*(GMB)[42]. Jest to duży korpus, który został wcześniej zanotowany semantycznie w sposób automatyczny. Głównym celem gier Wordrobe jest opracowanie złotego standardu anotacji, w postaci podkorpusu GMB. Wordrobe zostało zaprojektowane w postaci strony internetowej, na której znajduje się panel gracza. Anotacja tekstu odbywa się w formularzu, który niczym nie odbiega od profesjonalnej aplikacji do oznaczania metadanych przez ekspertów. Każda z gier przygotowana jest w przyjemnej oprawie graficznej i posiada czytelny interfejs (rys. 1) . Zarówno strona internetowa jak i korpus są w języku angielskim. Autorzy gier ograniczali ilość wyświetlanego tekstu do minimum, stosując zamiast niego czytelne dla użytkownika symbole.

Rysunek 1. Przykład gry z zestawu Wordrobe służącej do anotacji znaczenia słów



Źródło: www.wordrobe.org

Każda z gier opiera się na tych samych, jasno określonych zasadach. Należy wybrać jedną z dostępnych odpowiedzi, która według intuicji gracza jest poprawna. Jeżeli gracz nie chce udzielić odpowiedzi na jakieś pytanie, może je ominąć nie ponosząc przy tym żadnych konsekwencji. Jest to dobra praktyka, dzięki której gracz nie czuje się przymuszony do

39 Przykładem może być: Amazon Mechanical Turk: https://www.mturk.com, 30.12.2014.

40 L. von Ahn, L. Dabbish, *Designing Games with a Purpose*, „Communications of the Association for Computing Machinery" 2008, no. 51 (8), s. 63-64.

41 R. McCreadie, C. Macdonald, I. Ounis, *Identifying Top News Using Crowdsourcing*, „Information Retrieval" 2013, no. 16 (2), s. 191-192.

42 Strona korpusu The Groningen Meaning Bank: http://gmb.let.rug.nl/, 30.12.2014.

dokonania wyboru, co jednocześnie zmniejsza ilość niepoprawnych odpowiedzi. Sama rozgrywka jest losowa, ponieważ za każdym razem gracz otrzymuje inne zdanie do anotacji. Wyzwaniem, które stoi przed użytkownikiem, jest zebranie jak największej ilości punktów i osiągnięć przyznawanych za anotacje. Kolejnym elementem projektu panelu gracza jest tablica wyników. Zamieszczone są w niej aktualne wyniki z ostatnich pięćdziesięciu dni. Dzięki niej gracze mają wyraźny cel, mogą rywalizować i na bieżąco śledzić aktywności innych graczy. Autorzy projektu Wordrobe nie udostępniają szczegółów dotyczących formy panelu administracyjnego, jednakże na pewno takowy posiadają.

Etap trzeci: Realizacja projektu

Realizacja jest główną fazą każdego *crowdsourcingowego* projektu. W zależności od tego czy użytkownik ma wykonać niewielkie zadania na gotowej platformie, czy proces pozyskiwania danych jest bardziej skomplikowany, może ona trwać od kilku minut do kilku miesięcy lub nawet lat. W tej fazie wyróżnia się trzy podstawowe zadania:

- zarządzanie i kontrola obiegu zadań (ang. *task workflow and managment*),
- zarządzanie uczestnikami (ang. *contributor management*),
- kontrola jakości (ang. *quality control*).

Możliwość ciągłej kontroli projektu pozwala na wprowadzenie zmian w treści zadań, jeżeli okazały się problematyczne dla użytkowników. Jakość uzyskanych wyników pozwala natomiast ustalić, przez ilu użytkowników każde zadanie powinno być anotowane, aby uzyskany wynik był pewny.

Największym problemem tej fazy jest decentralizacja procesu pracy, ponieważ uczestnicy biorący udział w projekcie nie mają bezpośredniego kontaktu ze zleceniodawcą. Często problematyczne w tym przypadku jest zarządzanie zadaniami, odpowiednie przeszkolenie uczestników i umiejętność oceny jakości ich pracy. Aby uniknąć niepowodzeń w wymienionych sytuacjach, platforma *crowdsourcingowa* powinna być zaprojektowana w jak najprostszy do obsługi sposób. Wszystkie zadania, z którymi może spotkać się użytkownik powinny być wcześniej przedstawione w formie szkolenia. Szkolenie jest najczęściej osobnym elementem platformy, w którym użytkownik jest prowadzony krok po kroku przez przykładowe zadania. Zazwyczaj ma ono taką samą formę dla wszystkich osób i zostało przygotowane na podstawie ręcznie anotowanego fragmentu korpusu (złotego standardu). Dzięki temu, użytkownik w trakcie treningu jest informowany dokładnie, jakie błędy popełnia, a co wykonuje prawnie. Przypomina to mechanizm często wykorzystywany w grach komputerowych, w którym gracz przechodzi misję szkoleniową przed rozpoczęciem głównej rozgrywki.

Jednym z podstawowych zadań, wykonywanym podczas realizacji projektu, jest zaangażowanie jak największej ilości uczestników poprzez odpowiednie zachęcenie ich do pracy. Jest to szczególnie kłopotliwe w przypadku nieodpłatnych projektów, w których jedyną zachętą jest ciekawe spędzenie wolnego czasu.

W przypadku projektu Wordrobe, nie są udostępniane dane dotyczące zarządzania użytkownikami i zadaniami, które są przez nich wykonywane. Nie zostało przygotowane żadne szkolenie dla graczy, pokazujące jak należy rozwiązywać zadania. Prostota gier i otrzymywanie punktacji odpowiadającej poprawności udzielanej odpowiedzi okazały się być wystarczające. Wielu użytkowników wzięło udział w projekcie i przyczyniło się do uzyskania zadowalających wyników finalnej wersji korpusu, mimo że żaden z nich nie przechodził szkolenia.

W trakcie publikacji raportu dotyczącego projektu Wordrobe, na stronie internetowej zarejestrowanych było 962 graczy, którzy udzielili łącznie 41541 odpowiedzi we wszystkich siedmiu grach[43]. Pomimo, że taka liczba wydaje się być relatywnie duża, to biorąc pod uwagę wyniki dotyczące konkretnych zadań, ilość anotacji nie jest wystarczająca. Przykładowo, w grze polegającej na oznaczaniu znaczenia słowa (ang. *senses*) zebrano 5478 odpowiedzi dla 3121 różnych zdań[44]. Oznacza to, że zaledwie połowa zdań została zaanotowana co najmniej dwa razy (1,673 anotacji na zdanie). Aby móc zastosować mechanizmy ewaluacji w celu stworzenia złotego standardu, z otagowanych zdań wybrano tylko te, które zostały oznaczone przez większą ilość użytkowników.

Etap czwarty: Ewaluacja i agregacja wyników

Ostatnią fazą projektu *crowdsourcingowego* jest stworzenie dobrego jakościowo korpusu z zebranych anotacji. W celu wybrania anotacji, które znajdą się w finalnej wersji korpusu, używa się metod opartych na statystycznej ocenie zgodności pomiędzy anotatorami. Jeżeli w trakcie etapu realizacji projektu zostaną zebrane dodatkowe dane o użyt-

43 N. J. Venhuizen, V. Basilem K. Evang, J. Bos, dz. cyt., s. 403.
44 Tamże, s. 398.

kownikach, które dotyczą skuteczności i ilości przeprowadzonego tagowania, to mogą one zostać użyte do obliczenia owego stopnia zgodności (ang. *inter-annotator agreement*)[45]. Odpowiedzi udzielone przez bardziej doświadczonych w grze użytkowników, będą traktowane jako potencjalnie lepsze rozwiązania, niż odpowiedzi gracza, który dopiero zaczyna używać platformy.

Po dodaniu anotacji użytkowników do korpusu, następuje ostateczna ewaluacja jakości zebranych danych. Odbywa się ona przy użyciu klasycznych metryk znanych ze statystki: precyzji (ang. *precission*), czułości (ang. *recall*) i *f-measure*, będącym ważoną wypadkową dwóch poprzednich miar[46].

Głównym celem zestawu gier Wordrobe jest stworzenie złotego standardu anotacji, który charakteryzuje się wysoką miarą precyzji. Ewaluacje przeprowadzono na zbiorze testowym składającym się ze 115 zdań, w którym każde miało 6 możliwych odpowiedzi. Zbiór testowy został otagowany przez 4 ekspertów - tak stworzony korpus potraktowano, jako referencyjny. Następnie porównano go z korpusem otagowanym przez graczy. W celu wybrania odpowiedzi użytkowników zastosowano różne strategie. Przyjęte sposoby wyboru odpowiednich anotacji były zależne od wartości ustalonych progów procentowych. Progi określają, jaka minimalna liczba osób musi udzielić określonej odpowiedzi. W zestawieniu ewaluacji zaprezentowano wyniki dla każdej z przyjętych strategii (tab. 1).

Tabela 1. Statystyczne miary ewaluacji dla korpusu wynikowego projektu Wordrobe

Strategia wyboru	Precyzja	Czułość	F-measure
Najpopularniejsza odpowiedź	0,880	0,834	0,857
50% zgodności	0,882	0,782	0,829
70% zgodności	0,945	0,608	0,740
100% zgodności	0,975	0,347	0,512

Źródło: Tabela opracowana na podstawie N. J. Venhuizen, V. Basilem, K. Evang, J. Bos, dz. cyt., s. 400

Biorąc pod uwagę, że kluczową miarą złotego standardu jest precyzja, wyniki osiągnięte z wykorzystaniem platformy Wordrobe są bardzo dobre. Przy odpowiedniej strategii wyboru anotacji precyzja finalnego korpusu przekroczyła 97%. Kosztem wysokiego wyniku okazała się być niska czułość, co w praktyce oznacza niewielki rozmiar korpusu końcowego. Głównym problemem projektu okazała się mała liczba graczy. Projekt jest jednak cały czas kontynuowany, a dane zebrane za pośrednictwem Wordrobe sukcesywnie poprawiają zasoby korpusu GMB.

Podsumowanie

Istnienie problemów AI-zupełnych przyczyniło się do poszukiwań alternatywnych sposobów rozwiązywania problemów sztucznej inteligencji, nie opartych wyłącznie na pracy komputera. Od lat poszukuje się metody, która pozwoliłaby na komunikację z maszyną przy pomocy języka naturalnego. Pomimo że komunikacja jest dla ludzi czymś oczywistym, nadal nie istnieje sposób jej automatyzacji. Z uwagi na niepowodzenia związane z próbami użycia reguł do opisu języka, zaczęto używać statystycznych metod przetwarzania języka naturalnego.

Sukces podejścia statystycznego w rozwiązywaniu problemów NLP zależy od wielkości korpusu tekstowego. Dlatego przygotowanie rzetelnego zbioru danych jest kluczowym aspektem tworzenia statystycznego systemu sztucznej inteligencji. Ze względu na dużą liczbę specjalistów zaangażowanych w tworzenie takiego korpusu, proces ten jest czasochłonny i kosztowny. Część alternatywnych metod pozyskiwania korpusów językowych, opiera się na zachęceniu dużej liczby osób do jego tworzenia. Z takich metod korzysta się w ramach paradygmatu HC i *crowdsourcingu*. Jednym z najbardziej obiecujących podejść jest tworzenie gier skierowanych na cel. Nie wymagają one opłacania pracowników i można zaprojektować je w sposób, który będzie zachęcał do wykonywania zadań obliczeniowych.

Opisany powyżej zestaw gier Wordrobe wspomógł tworzenie rzetelnego korpusu językowego. Ten i wiele innych przykładów gier skierowanych na cel pokazuje, że z powodzeniem można wykorzystywać je do pozyskiwania zasobów językowych. W projekcie Wordrobe głównym problemem okazała się być niewielka liczba graczy, co skutkowało małym rozmiarem finalnego korpusu. Powodem tego może być forma zadań, które przypominają raczej zgrywalizowany proces

45 M. Sabou, K. Bontcheva, L. Derczynski, A. Scharl, dz. cyt., s. 864.

46 C. Goutte, E. Gaussier, *A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation*, „The European Conference on Information Retrieval" 2005, s. 346-347.

anotacji niż faktyczną grę. Gry skierowane na cel angażują użytkowników poprzez nagradzanie ich puntami, prowadzanie rankingów wyników i możliwość zdobywania osiągnięć. Możliwe jednak, że takie zabiegi nie są wystarczające, aby zmotywować do wielokrotnego powracania do rozgrywki. Biorąc pod uwagę liczbę osób, która codziennie spędza czas w grach online, potencjał gier skierowanych na cel wydaje się być nie w pełni wykorzystany. Problemu nie stanowi sposób zachęcania graczy do spędzania czasu w grach, ale zaoferowanie im rozgrywki konkurującej z popularnymi rozrywkami komputerowymi. Tworzenie eksperymentalnych narzędzi anotacji, takich jak Wordrobe, które wykorzystują niektóre narzędzia grywalizacji, udowodniło, że dzięki grom można pozyskać dane o wysokiej jakości. Jednakże, aby przyciągnąć większą ilość zainteresowanych należałoby zastosować inne mechanizmy, znane z gier online. Rozgrywkę mogłaby uatrakcyjnić ciekawa historia, logika gry, albo rywalizacja, czy kooperacja graczy, nie polegająca wyłącznie na zerkaniu na tabele z wynikami.

Bibliografia:

[1] von Ahn L., *Games With a Purpos*e, „Institute of Electrical and Electronics Engineers Computer Magazine" 2006.
[2] von Ahn L., Dabbish L., *Designing Games with a Purpose*, „Communications of the Association for Computing Machinery" 2008, no. 51 (8).
[3] Aker A., El-Haj M., Albakour M., Kruschwitz U., *Assessing Crowdsourcing Quality through Objective Tasks*, „Language Resources and Evaluation Conference" 2012.
[4] Bolshakov I. A., Gelbukh A., *Computational Linguistics: Models, Resources, Applications*, Mexico 2004.
[5] Goutte C., Gaussier E., *A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation*, „The European Conference on Information Retrieval" 2005.
[6] Grier D. A., *Foundational Issues in Human Computing and Crowdsourcing*, „Conference on Human Factors in Computing Systems" 2011.
[7] Grytczuk J., *Czy wszystko można obliczyć. Łagodne wprowadzenie do złożoności obliczeniowej*, [w:] Sysło M. M. (red.), *Homo Informaticus*, Warszawa 2012.
[8] Howe J., *The Rise of Crowdsourcing*, „Wired Magazine" 2006, no. 14 (6).
[9] Jones K. S., *Natural Language Processing: a Historical Review*, „Artificial Intelligence Review" 2001.
[10] McCreadie R., Macdonald C., Ounis I., *Identifying Top News Using Crowdsourcing*, „Information Retrieval" 2012, no. 16 (2).
[11] Nadkarni P. M., Ohno-Machado L., Chapman W. W., *Natural Language Processing: an Introduction*, „The Journal of the American Medical Informatics Association" 2011, no. 18 (5).
[12] Paroubek P., Chaudiron S., Hirschman L., *Principles of Evaluation in Natural Language Processing*, „French Association for Natural Language Processing" 2007, no. 48 (1).
[13] Quinn J., Bederson B. B., *Human Computation: a Survey and Taxonomy of a Growing Field*, „Conference on Human Factors in Computing Systems" 2011.
[14] Resnik P., Klavans J., *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, New Mexico 1994.
[15] Sabou M., Bontcheva K., Derczynski L., Scharl A., *Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines*, „Language Resources and Evaluation Conference" 2014.
[16] Samuel K., *The Gold Standard and the Origins of the Modern International Monetary System*, „Centre Études internationales et Mondialisation" 2003, no. 3(1).
[17] Surowiecki J., *The Wisdom of Crowds*, New York 2005.
[18] Sysło M. M., *Historia rachowania – ludzie, idee, maszyny. Historia mechanicznych kalkulatorów*, [w:] Sysło M. M. (red.), *Homo Informaticus*, Warszawa 2012.
[19] Szymański J., Duch W., *Annotating Words Using WordNet Semantic Glosses*, [w:] Huang T., Zeng Z., Li C., Leung C. S. (red.), *Neural Information Processing*, Doha 2012.
[20] Tkaczyk P., *Grywalizacja. Jak zastosować mechanizmy gier w działaniach marketingowych*, Gliwice 2012.
[21] Venhuizen N. J., Basilem K., Evang V., Bos J., *Gamification for Word Sense Labeling*, „IWCS" 2013.
[22] Yampolskiy R. V., *AI-Complete, AI-Hard, or AI-Easy – Classification of Problems in AI*, [w:] Visa S., Inoue A., Ralescu A. (red.), *Proceedings of the 23rd Midwest Artificial Intelligence and Cognitive Science Conference*, Cincinnati 2012.

Netografia:

[23] Adami A. G., *Automatic Speech Recognition: from the Beginning to Portuguese Language*, Materiały konferencyjne, International Conference on Computational Processing of Portuguese 2010, https://www.inf.pucrs.br/~propor2010/proceedings/tutorials/Adami.pdf, 31.12.2014.
[24] von Ahn L., *Human Computation*. (rozprawa doktorska), http://reports-archive.adm.cs.cmu.edu/anon/2005/CMU-CS-05-193.pdf, Carnegie Mellon University 2005, 31.12.2014.
[25] Amazon Mechanical Turk, https://www.mturk.com , 30.12.2014.
[26] Howe J., *Crowdsourcing: A Definition*, http://crowdsourcing.typepad.com, 31.12.2014.
[27] Gra Eterna, http://eternagame.org/, 30.12.2014.
[28] Gra Eyewire, http://eyewire.org, 30.12.2014.
[29] Gra Phylo, http://phylo.cs.mcgill.ca/, 30.12.2014.
[30] Gra Wordrobe, www.wordrobe.org, 30.12.2014.
[31] The Groningen Meaning Bank, http://gmb.let.rug.nl/, 30.12.2014.